Niladri Sekhar Dash · L. Ramamoorthy

Utility and Application of Language Corpora

Springer

Utility and Application of Language Corpora

Niladri Sekhar Dash · L. Ramamoorthy

Utility and Application of Language Corpora

Springer

Niladri Sekhar Dash
Linguistic Research Unit
Indian Statistical Institute
Kolkata, West Bengal, India

L. Ramamoorthy
Linguistic Data Consortium-Indian Languages
Central Institute of Indian Languages
Mysore, Karnataka, India

ISBN 978-981-13-1800-9 ISBN 978-981-13-1801-6 (eBook)
https://doi.org/10.1007/978-981-13-1801-6

Library of Congress Control Number: 2018949868

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

To
Prof. N. Deiva Sundaram

and

The crescent moon floating on the Marina beach

Preface

Between the desire
And the spasm
Between the potency
And the existence
Between the essence
And the descent
Falls the Shadow
For Thine is the Kingdom
T. S. Eliot: *The Hollow Man*

In recent years, there has been a notable upsurge in the use of language corpus as a dependable resource in all major domains of linguistics. The growing tendency for using language corpus as a source of authentic empirical evidence is triggered due to easy access to corpus data in a machine-readable form. Also, objective analysis of corpus data contributes in a major way to the understanding of the intricate form and function of language properties considered useful in various domains of linguistics. This change in approach reflects on the ideological and technological shifts the discipline has undergone over the last 60 years or more.

The ideological shift has focused on empirical language data as the most trusted linguistic evidence because our traditional intuitive assumptions are proved to be not much convincing for drawing reliable, conclusive observations about various aspects of a natural language on which principles and theories of linguistics can be formulated. This leads us to focus on empirical language data so that our understandings of languages and linguistics are reliable and trustworthy. The technological shift, on the other hand, has given us an opportunity to use a computer to collect language data in a new device, access it in different platforms, process it in different ways, analyze it from different perspectives, and utilize it in different domains. Thus, computer brings in a new shade in the world of linguistics to look at the language from newer angles to make our studies authentic and reliable.

In reality, the computer has opened before us many new ways of looking at the language data, interpreting it in a new light and utilizing it in various fields of linguistics. This was hardly possible before the use of the computer in linguistics. In the earlier years, scholars had to be content with handmade language data for their works because they had no system under their access by which they could assemble a large amount of language data from different spheres of language use within a short span of time, interpret them, and utilize them. Now the scenario is greatly changed. Now we can use a computer to collect language data of any size, type, and variety as well as analyze it to find out new examples, new information, and new evidence to furnish in our studies. The referential value of a corpus is increasing day by day because a corpus has the potential to contribute to new theory formulation as well as modifying the existing ones. Corpus-based studies are incorporating new insights to look into the cognitive areas of the human mind to understand the mysteries operating behind the cognitive process like receiving, processing, comprehending, and sharing linguistic signals (Winograd 1983: 18).

Nearly after six decades of corpus use in various fields of descriptive linguistics, applied linguistics, and language technology, it is clear that the utility of a corpus is not limited to language teaching and dictionary compilation. It is now an open resource which can be used in any domain of social sciences where language data is an integral part, although scholars more often were found to depend on it in the activities like language processing, machine learning, sentiment analysis, dictionary compilation, grammar writing, WordNet design, word-sense disambiguation, translation, language documentation, and other areas of linguistics (Dash and Arulmozi 2018: Chap. 14). Since the analysis of data in a language corpus supplies important authentic perspectives toward linguistic description and interpretation (Biber 1996), information retrieved from a corpus can also be used, besides the domains stated above, in several other linguistic works like diachronic lexical semantics, pragmatic analysis of texts, sociolinguistic studies, and discourse analysis (Leech and Fligelstone 1992). From an informal pilot survey, we have found that digital language corpora are used in more than 500 different types of development and research works covering almost all the major areas of linguistics and language technology (Dash and Chaudhuri 2003). The multipurpose application of language corpora is best visualized by Svartvik (1986):

> …lexicography, lexicology, syntax, semantics, word-formation, parsing, question-answer synthesis, software development, spelling checkers, speech synthesis and recognition, text-to-speech conversion, pragmatics, text linguistics, language teaching and learning, stylistics, machine translation, child language, psycholinguistics, sociolinguistics, theoretical linguistics, corpus clones in other languages such as Arabic and Spanish—well, even language and sex.

We can summarize the importance of a language corpus in the following manner. The relevance of corpus is understandable when we realize that the goal of a corpus is to perform the following tasks (Leech 1992):

(a) Focus on linguistic performance rather than competence,
(b) Focus on linguistic description rather than linguistic universals,
(c) Focus on quantitative and qualitative models of language,
(d) Focus on an empiricist, rather than a rationalist, view of scientific inquiry of a language.

Each of the arguments stated above contrasts with 'Chomskyan paradigm which has dominated much of linguistic thinking since the 1950' (Leech 1992: 107). The scope of corpus use is further expanded in the observations of Atkins et al. (1992), Leech and Fligelstone (1992), McEnery and Wilson (1996), Rundell (1996), Barlow (1996), Thomas and Short (1996), Biber (1996), Biber et al. (1998), Teubert (2000), Cheng (2011), Crawford and Csomay (2015), McEnery and Hardie (2011), Vandelanotte et al. (2014), Weisser (2015), Dash (2005), Dash and Arulmozi (2018), and many others. In essence, a language corpus is a valuable linguistic resource for all areas of descriptive linguistics, language technology, applied linguistics, discourse analysis, and cognitive linguistics. With the multidimensional use of language data, corpus linguistics now emerges as a new approach toward linguistics. It is also identified as a new way of studying and applying natural language using the techniques of computer science (Landau 2001: 277).

The appreciation for language corpus in linguistics is a new thing. It has been possible because the data and information obtained from a corpus have been of much use in developing corpus-based grammar and text materials for both first- and second-language learners. At the same time, the creation of a corpus, in both spoken and written forms, has been a new task of different orientation that has contributed in two different ways:

(a) It has developed a new league of scholars who are very much skilled in the generation of speech and text corpora as well as in processing them.
(b) It has developed a pool of data and information for catering the needs of different kinds of linguistic study and application.

The creation of this kind of ambience for the Indian languages is an urgent task today. Here our goal is to develop different types of the corpora in such a manner that they can contribute to new types of linguistic studies, research in written and spoken languages, and develop new technology that can serve the requirements of a multilingual and multicultural country like India. Although these corpora will be used primarily for linguistics and allied disciplines, these will also function as a source of data and information for various related disciplines. The corpora that we try to envisage will eventually be available online for each major and minor Indian languages and will be marked with genres, periods, times, and subjects—all being ready to mirror the discourse varieties that range over various types of linguistic interactions in which the members of the speech communities are engaged in.

From this observation, it becomes clear that in a multilingual and multicultural country like India, the functional potential of language corpus is multidimensional in various innovative works of descriptive linguistics, language technology, applied linguistics, cognitive linguistics, and discourse studies. Definitely, corpus-based

approach to linguistic research and development will open up many new avenues for the benefit of the people of the country. We can visualize that the corpora made from the Indian languages may be utilized in the following works: language description and interpretation, first- and second-language education, generation of terminology data banks, compiling lists of translational equivalents, compiling dictionaries of various types, studying traits of cultural difference across speech and language communities, analyzing courses of discourse and pragmatics in linguistic settings, analyzing dialogues and conversations in various settings and backgrounds, studying grammar and syntax of the Indian languages, developing grammar of different types in the Indian languages, developing systems and tools for understanding ambiguities, understanding the nature and texture of social psychology, analyzing systems and methods of language acquisition, devising tools and systems for language technology, designing texts and course materials for language education, exploring stylistic variations observed in different types of texts, studying language variation across regional and social territories, language revival and revitalization, translation across the Indian languages, and resource generation for serving language impairments.

Since a raw (i.e., non-normalized and non-annotated) corpus has limited utilities in analysis and description of a language or a variety, it has less value in works of language description, application, and technology. Therefore, there arises an urgent need for generating normalized and annotated corpora in the Indian languages for both spoken and written varieties. Since there is hardly any Indian language corpus that is tagged at various levels (i.e., orthographic, grammatical, syntactic, semantic, discourse, anaphoric, etymological, and figurative) of a text, there is an urgent need for generating normalized and annotated corpora, which will be used to contribute toward devising advanced resources and tools for the works of applied linguistics and language technology. In order to achieve this goal, we need to design elaborate tagsets, which will encompass all linguistic properties of written and spoken texts with a proper focus on sounds, segments, characters, words, morphemes, compounds, collocations, multiword units, phrases, sentences, idioms, proverbs, and other linguistic properties. There is also a need for describing different methods of text annotation for the corpora of Indian languages based on which subsequent works of corpus processing and utilization can be fruitfully executed (all these issues will be discussed in some details in the next volume).

Besides tagging these linguistic elements, properties, and information visible in a corpus, we also need a corpus tagged with various invisible linguistic information such as meaning, anaphora, discourse, and pragmatics so that these corpora become useful and accessible for various applied linguistic works relating to dictionary compilation, language teaching, word-sense disambiguation, machine translation, etc. In this context, the tools and techniques that we require for all the Indian language corpora are the following: frequency counting, lexical sorting, sentence breaking, lexical decomposition, compound decomposition, spelling checking, word concordance, lexical collocation, keyword searching, local word grouping, lemmatization, morphological processing, morphological generation, grammatical annotation, sentence annotation, multiword extraction, sentence parsing, Sandhi

splitting, speech-to-text conversion, and transliteration. Although these tools and techniques are absolutely necessary for the Indian language corpora, it is not so easy to develop these tools and resources without the support of actual Indian language corpora and active involvement of a large group of skilled people endowed with advanced knowledge of Indian languages, corpus linguistics, and computer science. This appears to be a Himalayan task before we can make a real breakthrough in the areas of corpus linguistics and language technology for the Indian languages. We have addressed some of the issues and needs mentioned above in this volume.

The present book is an outcome of our corpus-based studies on the Indian languages. In this book, we have tried to address some of the issues of using data and information from the corpus in some basic areas of language application. We have tried to show how language corpus may be utilized for developing quality language teaching textbooks, course books, and reference materials; compiling usage-based dictionaries; generating database for translational equivalents; deciphering contextual information for understanding sense variations in words; understanding unique linguistic properties and features of dialects; and developing useful systems, tools, and resources for translation. In essence, with close reference to the corpus, in this book, we have tried to perform the role of a harbinger to make people aware of functional and referential benefits of the corpus in various works of descriptive linguistics, applied linguistics, and language technology. The academic relevance of this book may be attested in its direct focus on the application of language data and its sincere appeal for redirecting the focus of linguists toward this new method of language study for the benefit of the entire discipline as well as its sister disciplines. Following are the notable contributions of this book:

[1] It discusses the use of corpus in several important areas of linguistics, namely language teaching, dictionary making, dialect study, word-sense disambiguation, TermBank compilation, machine translation, lexicology, sociolinguistics, lexical semantics, psycholinguistics, stylistics.
[2] The observation, information, and arguments furnished in this book are based on the analysis of language corpus of various types.
[3] From the perspective of application of corpus in the areas stated above, we have tried to show how linguistic information and examples obtained from a corpus can contribute toward the growth, maturity, and advancement of linguistics.
[4] Discussions presented in this book are primarily built on linguistic resources collected from the real use of language in normal life. These are authentic and reliable than any experimental results and intuitive speculations.
[5] The research works presented in this book focus on the qualitative and functional interpretation of corpus data to understand the intricacies noted in the internal and external textures of a natural language.

[6] The topics discussed in this book have strong academic and functional relevance in the general domains of corpus linguistics, applied linguistics, language technology, cognitive linguistics, computational linguistics, natural language processing, and mainstream linguistics.
[7] The book tries to show how new findings obtained from a corpus become useful to substantiate, validate, or refute previously made observations and hypotheses about language properties.
[8] The book searches answers to those queries of linguistics and language technology which are relevant and useful for future studies and research in corpus linguistics, language technology, and applied linguistics.
[9] The book tries to find answers to the questions indirectly related to the cognitive, functional, and referential relevance of corpus in the main areas of linguistics and language technology, which has been haunting scholars for the last fifty years.
[10] The book is enriched with reference to recent works carried out in several advanced languages in various parts of the world. It will help the readers to know how and where novel approaches are used and how these are making valuable improvements over traditional systems, models, and techniques normally used in various domains of linguistics.

Since the book is highly referential in approach and analysis, it is characteristically suitable to be used as a course-cum-text book at the undergraduate and the postgraduate levels. It can also be used as a reference book for language teachers of first- and second-language teaching, researchers working in various areas of linguistics, and people engaged in the development of tools and systems of language technology. Also, people working in corpus linguistics, computational linguistics, natural language processing, applied linguistics, cognitive linguistics, lexicography, discourse analysis, lexicology, field linguistics, language documentation, translation, etc., will find this book equally useful for relevant information, observation, and interpretation.

Kolkata, India — Niladri Sekhar Dash
Mysore, India — L. Ramamoorthy
March 2018

References

Atkins, S., J. Clear, and N. Ostler. 1992. Corpus design criteria. *Literary and Linguistic Computing*. 7, no. 1: 1–16.

Barlow, M. 1996. Corpora for theory and practice. *International Journal of Corpus Linguistics.* 1, no. 1: 1–38.

Biber, D. 1996. Investigating language use through corpus-based analyses of association patterns. *International Journal of Corpus Linguistics*. 1, no. 2: 171–198.

Biber, D., S. Conrad, and R. Reppen. 1998. *Corpus linguistics: Investigating language structure and use.* Cambridge: Cambridge University Press.
Cheng, W. 2011. *Exploring corpus linguistics: Language in action.* London: Routledge.
Crawford, W. and E. Csomay. 2015. *Doing corpus linguistics*. London: Routledge.
Dash, N.S. 2005. *Corpus linguistics and language technology: With reference to Indian languages.* New Delhi: Mittal Publications.
Dash, N.S. and S. Arulmozi. 2018. *History, features, and typology of language corpora.* Singapore: Springer.
Dash, N.S., and B.B. Chaudhuri. 2003. Relevance of language corpora in language research and application. *International Journal of Dravidian Linguistics.* 32, no. 2: 101–122.
Landau, S.I. 2001. *Dictionaries: The Art and Craft of Lexicography.* Revised Second Edition. Cambridge: Cambridge University Press.
Leech, G. 1992. *Corpora and theories of linguistic performance*. Edited by J. Svartvik, *Directions in corpus linguistics: Proceedings of nobel symposium 82-Stockholm.* Berlin: Mouton De Gruyter. 105–122.
Leech, G. and S. Fligelstone. 1992. *Computers and corpus analysis*. Edited by C.S. Butler, *Computers and written texts.* Oxford: Blackwell Publishers. 115–140.
McEnery, T. and A. Hardie. 2011. *Corpus linguistics: Method, theory, and practice.* Cambridge: Cambridge University Press.
McEnery, T. and A. Wilson. 1996. *Corpus linguistics.* Edinburgh: Edinburgh University Press.
Rundell, M. 1996. The corpus of the future and the future of the corpus. Invited talk delivered at a special conference on the new trends in reference science at Exeter, UK (a hand out).
Svartvik, J. 1986. For Nelson Francis. *International Computer Archive of Modern English News*. no. 10: 8–9.
Teubert, W. 2000. Corpus linguistics—a partisan view. *International Journal of Corpus Linguistics.* 4, no. 1: 1–16.
Thomas, J. and M. Short. ed. 1996. *Using corpora for language research: Studies in honor of Geoffrey Leech*. London and New York: Addison Wesley Longman.
Vandelanotte, L., K. Davidse, C. Gentens, and D. Kimps. 2014. *Recent advances in corpus linguistics: Developing and exploiting corpora*. Amsterdam: Rodopi.
Weisser, M. 2015. *Practical corpus linguistics: An introduction to corpus-based language analysis*. London: Wiley-Blackwell.
Winograd, T. 1983. *Language as a cognitive process.* Vol. I. Mass.: Addison-Wesley.

Acknowledgements

We sincerely thank our seniors, peers, and juniors who have helped us in different capacities to shape up our observations and ideas in the form of this book. We also acknowledge those known and unknown scholars from whom we have tried to assimilate insights and information to formulate our concepts furnished in this book. We humbly thank those unknown reviewers who have suggested necessary corrections and modifications for the improvement of content and quality of the book. We sincerely appreciate their wise and insightful comments for the betterment of the book.

We acknowledge the support and encouragement we have received from our parents, teachers, colleagues, and friends for writing this book. Particularly, we would like to mention the name of Ms. Shinjini Chatterjee and Ms. Priya Vyas who have been persistently encouraging us to shape up our thoughts to write this volume. This book would not have been possible without their continuous encouragement and support.

Niladri happily expresses his sincere thanks to Soma, Shrotriya, and Somaditya for their perennial emotional support and encouragement they extended during the course of writing this book. They have always been with him to boost up his morale during odd circumstances and adverse situations.

We shall consider our efforts are amply rewarded if people interested in corpus linguistics and sister domains find this book useful for their academic and non-academic endeavors.

March 2017

Niladri Sekhar Dash
L. Ramamoorthy

Contents

About the Authors

Dr. Niladri Sekhar Dash works as Associate Professor in *Linguistic Research Unit, Indian Statistical Institute, Kolkata* (The Institute of National Importance, Government of India). For the last 25 years, he has been working in the areas of *corpus linguistics, language technology, language documentation and digitization, computational lexicography, computer-assisted language teaching,* and *manual and machine translation*. To his credit, he has published 16 research monographs and more than 250 research papers in peer-reviewed international and national journals, anthologies, and conference proceedings. As an invited speaker, he has delivered talks at more than 40 universities and institutes in India and abroad. He acts as Research Advisor for several multinational organizations (e.g., Zi Corporation (Canada), Mobile Net (Sweden), Taylor & Francis (England and USA), Cognizant (India), Brahmin Ltd. (India), Mihup (India), Beno Translation (USA), Oxford University Press (UK), Reve Systems Ltd. (Bangladesh)), which have been working on language technology, natural language processing, and computational lexicography. He is Principal Investigator for three language technology projects funded by DeitY, MeitY, Government of India, and ISI—Kolkata. He is Editor-in-Chief of *Journal of Advanced Linguistic Studies*—a peer-reviewed international journal of linguistics enlisted in UGC journal list (2017)—and Editorial Board Member of six international journals. He is a member of several linguistic associations across the world and a regular Ph.D. thesis evaluator of several Indian universities. Recently, he is awarded *British Academy Visiting Fellowship 2018*—a highly prestigious and coveted fellowship bestowed by the *British Academy, UK.* At present, he is working on digital pronunciation dictionary, computer-assisted language teaching, digital lexical profile, contextualized word-sense cognition, parallel translation corpus, bilingual lexical database, POS tagging, language documentation and digitization, and human and machine translation, etc. Details of him are at https://sites.google.com/site/nsdashisi/home/.

Dr. L. Ramamoorthy is currently working in the *Central Institute of Indian Languages*, a subordinate office of the *Ministry of Human Resource Development*, Government of India. He is heading the *Centre for Corpus Linguistics* under which

the *Linguistic Data Consortium for Indian Languages* (LDC-IL) is functioning. LDC-IL developed quality-annotated corpus in terms of text and speech for all the scheduled languages of India. He conducted many training programs throughout the country for developing manpower in language technology. His areas of specialization include sociolinguistics, language planning, and education technology. As the head of the program on education technology, he supervised the production of more than 200 visual episodes on language, literature, and culture, which can be used as supplementary materials for language teaching. He also supervised in developing online teaching materials for Urdu, Telugu, and Manipuri. He was also heading the *Scheme for Protection and Preservation of Endangered Languages* (SPPEL) through which the Institute planned to document endangered languages in India. He supervised the preparation of handbook and guidelines for documentation and trained many scholars in the documentation. He was Director-in-Charge of the *Pondicherry Institute of Linguistics and Culture*, an autonomous body under the Government of Pondicherry. He has published seven books on various topics like language modernization, language loyalty, and purism. He also edited eight books and published many articles in reputed journals. Till date, 15 candidates have been awarded Ph.D. degree under his supervision.

Abbreviations

AE	American English
AFC	Asymmetrical Frequency Characteristics
AIMS	Access of Information from Multiple Sources
ANC	American National Corpus
ANOVA	Analysis of Variance
ASCII	American Standard Code for Information Interchange
BBC	British Broadcasting Corporation
BE	British English
BIS	Bureau of Indian Standards
BNC	British National Corpus
BoE	Bank of English
BPTC	Bilingual Parallel Translation Corpus
CALT	Computer-Assisted Language Teaching
CBA	Corpus-Based Approach
CBMT	Corpus-Based Machine Translation
CCH	Common Core Hypothesis
C-DAC	Center for Development of Advanced Computing
CIIL	Central Institute of Indian Languages
COLT	Corpus of London Teenagers
DDL	Data-Driven Learning
DOS	Disk Operating System
EBMT	Example-Based Machine Translation
ELC	English Language Corpus
ELT	English Language Teaching
FDT	Free Discourse Text
FLT	First-Language Teaching
GC	Global Context
GIST	Graphics and Intelligence-based Script Technology
GRD	General Reference Dictionary
ICAME	International Computer Archive of Modern English

ICE	International Corpus of English
ICLE	International Corpus of Learner English
IE	Indian English
ILCI	Indian Language Corpora Initiative
ILPOST	Indian Languages Part-of-Speech Tagging
IPA	International Phonetic Alphabet
ISCII	Indian Standard Code for Information Interchange
ISI	Indian Statistical Institute
KBA	Knowledge-Based Approach
KCIE	Kolhapur Corpus of Indian English
KWIC	Key-Word-In-Context
LC	Local Context
LDC	Linguistic Data Consortium
LDC-IL	Linguistic Data Consortium for Indian Languages
LLC	Lancaster-Lund Corpus
LOB	Lancaster-Oslo-Bergen
LT	Language Technology
LWG	Local Word Grouping
MAT	Machine-Aided Translation
MCIT	Ministry of Communication and Information Technology
MHRD	Ministry of Human Resource Development
MIS	Mutual Information Scale
MRD	Machine-Readable Dictionary
MT	Machine Translation
NLP	Natural Language Processing
NW	Neighboring Word
OCR	Optical Character Recognition
OS	Operating System
OTA	Oxford Text Archive
PD	Perceptual Dialectology
POS	Part-of-Speech
RTF	Rich Text Format
RWE	Real Word Error
SBMT	Statistics-Based Machine Translation
SC	Sentential Context
SL	Source Language
SLT	Second-Language Teaching
SP	Script Processor
ST	Scientific Term
TC	Topical Context
TC	Translation Corpus
TDIL	Technology Development for the Indian Languages
TE	Translational Equivalent
TL	Target Language

TT	Technical Term
TU	Translation Unit
TW	Target Word
UGC	University Grant Commission
WWW	World Wide Web

List of Figures

Chapter 1
Issues in Text Corpus Generation

Abstract In this chapter, we shall briefly discuss some of the basic issues that are directly linked with text corpus generation in digital form with the involvement of computer in the process. The act of corpus generation asks for consideration of various linguistic and statistical issues and factors which eventually control the entire process of corpus generation. Factors like size of a corpus, choice of text documents, collection of text documents, selection of text samples, sorting of text materials, manner of page sampling and selection, determination of target corpus users, manner of data input, methods of corpus cleaning, management of corpus files are immediate issues that demand utmost attention in corpus generation. Most of these issues are important in the context of text corpus generation not only for advanced languages like English and Spanish but also are of greater importance for poorly resourced languages used in less advanced countries. We shall discuss all these issues in this chapter with reference to some of the Indian languages.

Keywords Size of corpus · Text representation · Determination of time span
Selection of documents · Selection of newspapers · Selection of books
Selection of writers · Determination of target users

1.1 Introduction

There are many issues involved in the generation of a corpus in digital form with texts taken from written sources. It asks for serious consideration of various linguistic, extralinguistic, and statistical factors that are directly linked to the process of corpus generation. Issues like size of a corpus, choice of text documents, collection of text documents (e.g., books, journals, newspapers, magazines, periodicals), selection of text samples, sorting of text materials, manner of page selection (e.g., random, regular, selective), determination of target users, manner of data input, methods of corpus cleaning, management of corpus files demand good care and attention from the corpus designers for the successful creation of a corpus (McEnery and Hardie 2011; Crawford and Csomay 2015).

N. S. Dash and L. Ramamoorthy, *Utility and Application of Language Corpora*,
https://doi.org/10.1007/978-981-13-1801-6_1

At the time of creating a corpus, all these issues are, however, not equally relevant to all types of corpus irrespective of languages. It is observed that some of the issues as proposed in Atkins et al. (1992) may be redundant for many of the less advanced languages including that of Indian and other South Asian languages. On the contrary, there are some issues, which are of high referential relevance for the Indian and the South Asian languages and are hardly addressed and probed into. That means it is possible to classify the corpus generation issues into three broad types based on the status of a language:

(a) Issues relevant for an advanced language,
(b) Issues relevant for a less advanced language, and
(c) Issues relevant for both types of languages.

In this chapter, we shall try to address most of the issues that are directly relevant for the less advanced languages used in India and South Asian countries. And, most of these issues are discussed in the following subsections of this chapter with direct reference to the Indian text corpus developed in the Technology Development for the Indian Languages (TDIL) project executed across the country during 1992–1995. In our view, the issues that will be discussed here are relevant not only for the Indian languages, but can also be relevant for other less advanced languages used across the world.

In Sect. 1.2, we shall express our views about the size of a general corpus; in Sect. 1.3, we shall focus on the issue of representation of text types in a corpus; in Sect. 1.4, we shall discuss the importance of determining time span for a corpus; in Sect. 1.5, we shall address the method of selection of text documents; in Sect. 1.6, we shall discuss the process of selection of newspapers texts; in Sect. 1.7, we shall describe the process of selection of books; in Sect. 1.8, we shall address the process of selection of writers of texts; and in Sect. 1.9, we shall focus on the importance of selection of target users for a corpus.

1.2 Size of a Corpus

The application of scientific data sampling technique in the act of corpus generation is a highly useful safeguard in the process of determining beforehand the hierarchical structure of use of language by the members of that language community. With the application of this method, we can clearly refer to the different genres and text types from where we can collect the required amount of data. For example, when we try to develop a written text corpus, we have to focus on collecting the adequate amount of samples from all kinds of texts that are produced, printed, published, and procured by the people in a language. The problem is that in less advanced languages, the numbers of texts belonging to these categories are not evenly distributed. That means there is an anti-pyramidical distribution of texts in less advanced languages where written texts are mostly produced in the traditional manner without much use of digital technology (Fig. 1.1).

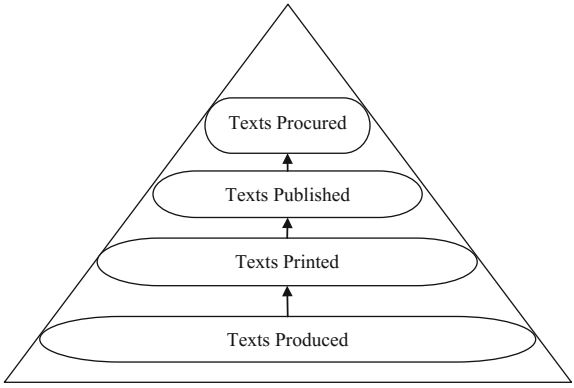

Fig. 1.1 Texts produced, printed, published, and procured by people

In this situation what is available to us only a few texts produced in printed form and we have no other option but to use these texts to generate a digital corpus. In most cases, our attention is focused on the easily accessible texts found in newspapers, books, journals, magazines, and other forms of printed sources (Cheng 2011). And most cases, these printed sources can provide us texts relating news events, fictions, stories, folk tales, legal statutes, scientific writings, social science texts, technical reports, government circulars, public notices, and so on. These texts are produced for general reading and reference as well as for other academic activities by the members of the speech community. The producers of these texts have never visualized in the line that such texts can have long-term applicational relevance if these were rendered into electronic version in the form of digital text corpora (Vandelanotte et al. 2014). This implies that the generation of a text corpus in less advanced languages, where digital texts are hardly available, is a very hard task. The target can be achieved if only a well-planned scheme of work is envisaged and implemented with due importance.

On the other hand, we can think of generating a speech corpus in a more simplified manner with representative samples of spoken texts collected from various speech events that occur at different times and places in the daily course of life of the speech communities. Collection of speech data is not a difficult task, but rendering these data in the form of a speech corpus is, however, a highly complicated task that may invoke complex processes like transcription and annotation before the speech data is appropriately marked as a 'speech corpus.' The process of sampling of speech data in a speech corpus is a statistical problem where the proper representation of speech samples has to be determined based on the percentage of use of various types of speech at different linguistic events by the language users (McEnery et al. 2005). This issue, however, is not elaborated further in this chapter, as we like to focus more on the size of a text corpus for less advanced languages.

The issue of size of a text corpus is mostly related to the number of text samples included in it. Theoretically, it can be determined based on the following two parameters:

(1) Number of sentences in each text sample and
(2) Number of words in each sentence.

In actuality, it is the total number of words that eventually determines the size of a corpus. Word is given more importance because the number of words in a sentence may vary based on the structure of a sentence. A sentence can have one or two words, while another sentence can have more than hundred words in it. It is, therefore, better to consider the word as a counting unit in determining the size of a corpus. There is nothing objectionable if anybody wants to determine the size of a corpus based on the number of sentences included in it. In general, a corpus, which includes more number of words, is considered bigger than a corpus, which includes less number of words.

Since size is considered an important issue in corpus compilation as well as in corpus-based language study, we must make a corpus as large as possible with adequate collection of texts from the language used in normal situations. The general observation is that a corpus containing 1 million words may be adequate for specific linguistic studies and investigation (Sinclair 1991: 20), but for a reliable description of a language as a whole, we perhaps need a corpus of at least 10 million words. One may, however, argue that even a corpus with 10 million words is not enough if the corpus is unidirectional. In that case, scanty information may be available for most of the words compiled in the word list. In the new millennium, probably, a corpus of 100 million words is the rule of the game. Given below is a list of some of the most prominent corpora of some of the most advanced languages of the world (Apresjan et al. 2006) (Table 1.1). It will clearly show how much data they contain and how big they are.

At the beginning of text corpus generation, the first question that comes to our mind is its total number of words. How big should a corpus be? The question, however, is related to the issues of making a text corpus properly representative and adequately balanced to make it maximally reliable and authentic source of data. It is also related to the number of 'tokens' and the number of 'types' included in the corpus. Also, it calls for the decision of how many text categories are to be kept in the corpus, how many text samples to be taken from each category, and how many words should be there in each text sample. All these may be applied faithfully in case of a large corpus which enjoys certain advantages, but may not be used properly for a small corpus which is usually deprived of many features of a general large corpus.

At the initial stage, when the work of corpus generation starts, the size of a corpus can be an important issue, since the process of word collection is a rigorous and tiresome task, which is usually carried out manually since digital texts materials are scanty and rarely available. Also, the idea that the corpus developed in this manner may be within the manageable range of manual analysis. Today, however, we have large and powerful computers where we can collect, store, manage, process, and analyze millions of words with high speed and optimum accuracy. Therefore, size is

Table 1.1 Some prominent and large language corpora of the world

No.	Name	Language	Word count
1	American National Corpus	American English	100 million+
2	Bank of English	British English	650 million+
3	British National Corpus	British English	500 million+
4	Corpus of Contemporary American English	American English	425 million+
5	Oxford English Corpus	Many Englishes	2.1 billion+
6	Croatian Language Corpus	Croatian	100 million+
7	Russian National Corpus	Russian	600 million+
8	Slovenian National Corpus	Slovenian	621 million+
9	National Corpus of Polish	Written Polish	1 billion+
10	German Reference Corpus	German	4 billion+
11	Spanish Historical Corpus	Spanish	100 million+
12	Spanish Dialect Corpus	Spanish	2 billion+
13	Chinese Internet Corpus	Chinese	280 million+
14	Chinese Business Corpus	Chinese	30 million+
15	Contemporary Written Japanese	Japanese	110 million+

not an important issue in the present state of corpus generation (Gries 2016). What we understand is that although size is not a good guarantee for proper text representation in a corpus, it is one of those vital safeguards that can shield a corpus from being skewed and non-representative of a language.

Although size affects validity and reliability, a corpus however big it may be, it is nothing but a small sample of all the language varieties that are produced by the users of that language (Kennedy 1998: 66). This signifies that within the frame of qualitative analysis of language data, size may become almost irrelevant. In contrast to the large-scale text corpora produced in English, Spanish, and many other advanced languages (Table 1.1), the size of the corpora produced in Indian and some South Asian languages is really small. Even then, the findings elicited from these small corpora do not vary much from that of the large corpora. For instance, the TDIL corpus that contains approximately 3 million words for each of the Indian 'national' languages, information derived from these corpora do fit, more or less, to the general linguistic features of the languages. In spite of this, we argue that we should venture in the direction of generating larger multidimensional corpora for most of the Indian languages and their regional varieties so that these corpora can be adequately representative of the languages both in data and formation.

1.3 Representation of Text Types

The question of size becomes irrelevant in the context of representation of text samples in a corpus. A corpus may be very large in size, which, however, does not guarantee that it is properly balanced to represent all possible varieties of use of a language. A large collection of text samples is not necessarily a corpus until and unless it does possess the feature of 'generalization' of language properties. That means a corpus can be true 'representative' only when the findings retrieved from its analysis can be generalized to the language as a whole or to a specified part of it (Leech 1991). Therefore, along with focussing on 'quantity of data,' we should equally emphasize on 'variety of data' so that the text samples from all possible domains of language use are proportionately represented within a corpus to make it maximally representative of the language under consideration.

To achieve proper representativeness, the overall size of a corpus may be set against the diversity of sources of text samples because within available text categories, the greater the number of individual samples, the greater is the amount of representation as well as greater is the reliability of analysis of linguistic variables (Kennedy 1998: 68). This settles the issue of proper representation of text samples within a corpus.

There are some important factors relating to balance and text representation within a corpus. It is noted that even a corpus of 100 million words can be considered small in size when it is looked into from the perspective of a total collection of texts from which a corpus is sampled (Weisser 2015). In fact, differences in the content of a particular type of text can influence subsequent linguistic analysis since the topic of a text plays a significant role in drawing inferences. Therefore, for an initial sampling of texts, it is better to use a broad range of objectively defined documents or text types as its main organizing principle. As a safeguard, we may use the following probabilistic approaches for the selection of text samples for a corpus (Summers 1991: 5):

(1) Apply an approach based on academic merit or influence of writers.
(2) Apply a method of random selection of text samples.
(3) Emphasize on currency or extent to which the texts are read.
(4) Use the method of subjective judgment of 'typicalness' of texts.
(5) Rely on the availability of texts in archives or other sources.
(6) Consider a demographic sampling of readers based on reading habits.
(7) Make empirical adjustments on texts to meet linguistic specifications.
(8) Justify purposes of investigators at the time of corpus building.

The most sensible and pragmatic approach is the one in which we try to combine all these criteria in a systematic way and where we can have data from a wide range of sources and text types with due emphasis on their 'currency,' 'typicalness,' and 'influence.'

The method of random text sampling is a powerful safeguard to protect a corpus from being skewed and non-representative. It is a standard technique widely used in many areas of natural and social sciences. However, we have to determine the

Table 1.2 List of text types included in the TDIL Indian language corpus

No	Text types	Year span	No. of words	%-age
1	Mass media	1981–1990	9,00,000	30
2	Creative writing	1981–1990	4,50,000	15
3	Natural science	1981–1990	3,00,000	10
4	Social science	1981–1990	3,00,000	10
5	Engineering and technology	1981–1990	3,00,000	10
6	Fine arts	1981–1990	1,50,000	5
7	Medical science	1981–1990	1,50,000	5
8	Commerce and industry	1981–1990	1,50,000	5
9	Legal and administration	1981–1990	1,50,000	5
10	Others	1981–1990	1,50,000	5
	Total		30,00,000	100

kind of language we want to study before we define the sampling procedures for it (Biber 1993). A suitable way to do this is to use bibliographical indexes available in a language. This is exactly what we have done for the Indian language corpora developed in the TDIL project. With marginal deviation from the method adopted for the *Brown Corpus*, we have used some (not all) major books and periodicals published in a particular area and specific year to include in the corpus (Table 1.2).

The number of words collected in the TDIL corpus is relatively small in comparison with the collection of words stored in the *British National Corpus*, the *American National Corpus*, the *Bank of English*, and others. However, we are in a strong position to claim that these text samples are well represented since the documents that are taken for inclusion are collected from all domains we found in printed form. The frequency of published documents used for the purpose of corpus development is presented in Table 1.2 to show that the majority of Indian people usually read newspapers and magazines more often than published materials belonging to different subject areas and disciplines (Dash 2001).

1.4 Determination of Time Span

A living language has a unique quality to change with time. Therefore, determination of a particular time span becomes essential at the time of text corpus generation. Once a time span is fixed, corpus users know that the language of a particular time period is represented in the corpus. It has another advantage for the corpus users who are interested to study the change of language across time. They chronologically arrange several synchronic corpora of particular text types to develop a diachronic corpus. For instance, consider that we have a few synchronic corpora of Indian languages, each one of which represents a decade of the twentieth century. By arranging all

these synchronic corpora in simple chronological order, we can produce a diachronic corpus of the twentieth century to track the language used through the century. Thus, a diachronic corpus becomes a valuable resource to study the chronological development and changes of the linguistic features over time.

For the purpose of generating the TDIL corpus, we selected a time period, which spanned from the year 1981 to 1990. This indicates that the text samples are collected from books, magazines, newspapers, reports, and other documents, which are printed and published within this time span. People may, however, raise a question regarding the relevance of selection of this particular time span. They can ask if this time span shows any special feature of the language that is not found in the language of other periods. The answer lies in technical reasons, common sense, and general knowledge rather than in linguistics. When we started the work of corpus generation in 1991, we faced severe difficulty in the act of collecting printed text materials published nearly 20 or 30 years ago. Although some books and journals were available, other printed text materials, particularly newspapers, government circulars, public reports, legal texts, little magazines, etc., were not readily available. Therefore, to overcome the difficulties of procuring old text materials as well as to make the project successful, we decided to divert our attention toward the text materials which were published in the previous decade. This solved some of the bottlenecks of the TDIL project.

Even then, contrary to our expectations, numerous unprecedented problems cropped up once we started the actual task of text collection. Collection of books published within 1981–1990 was not much tough. We were able to collect these materials from libraries of schools, colleges, and universities, as well as from personal collections. We tasted similar success in case of journal papers, which we mostly collected from personal collections and institutional libraries.

The task of collecting newspapers, magazines, and periodicals published 10 years ago was almost an impossible mission. No newspaper house cooperated with us. While some houses were skeptical about the relevance of the project, others asked very high price for old papers and documents. On the other hand, central, state, and public libraries were not willing to give newspapers for data collection. As a result, the task of collecting newspapers, magazines, and periodicals was hampered to a great extent, which affected the overall balance and composition of the TDIL corpus (Dash 2007).

The text materials that we collected from the personal collection were a good safeguard in the whole enterprise. However, we also faced some problems there and these were tackled with careful manipulation of text documents. For instance, there was no consistency of text types in case of personal collection as this kind of collection is usually controlled by an individual's preference, occupation, choice, and other personal factors. What we noted is that if we found a copy of a newspaper of a particular year (say, 1982), we invariably failed to procure a copy of that particular newspaper of the previous (i.e., 1981) or the next year (i.e., 1983). Most often, the solution to this problem was found in the collection of scrap paper collectors who had supplied many newspapers and magazines which were not found in the personal collection.

There was another crucial problem with regard to the selection of time span in case of text document collection, particularly for printed books. It was found that there were a large number of books, which were first published before the scheduled time span, and again were reprinted within the selected time span. The question was whether we should collect texts from these books as the first publication date of these books was much before the time span selected for the project. We had to decide carefully if such text materials could be fit to be considered for inclusion in the TDIL corpus.

The final decision, however, was laid with the corpus designers. Since we found that most of the texts were quite relevant in the then state of the language, we included them in the list. This kind of challenge can be entertained in case of synchronic corpus where texts are meant to be obtained from a specific time span to analyze time-stamped features of a language. In case of the diachronic corpus, such restriction does not hold any relevance as a diachronic corpus, by virtue of its nature and composition, and is entitled to include all types of text obtained from text materials published across several years.

1.5 Selection of Text Documents

Selection of text documents and collection of data from these text documents are two complex methods that require careful analysis and implementation by corpus designers. For ease and accuracy in data sampling, there are some well-known statistical methods which may be used (Barnbrook 1998). The first thing, however, is to identify the types of books and journals from where texts can be procured for the corpus.

In case of a general corpus, this is less troublesome since a general corpus can take data from all kinds of text documents. Here, the emphasis is given more on the amount of text data than on the types of the text sample. Following a simple method of text representation, samples from all text types may be included here without much consideration of the types of text. On the other hand, if a corpus is a 'special corpus,' then we have to be much careful in the selection of text types; else, the corpus will fail to highlight the special feature of a language for which it is made. Since the TDIL corpus is a general multidisciplinary monolingual general corpus, there is less trouble in the selection of documents for data collection. Therefore, anything printed and published within the scheduled time period is worth selection for retrieving the fixed amount of text data for the corpus.

The general argument of the corpus designers in this context was that each year should have an equal amount of text representation. That means no year would have a larger amount of data than its preceding or succeeding year. This would help us in maintaining the overall balance of the TDIL corpus. The statistics that have been given below (Table 1.3) provide a general idea of how the total number of words was collected from various text documents spreading over the years.

Table 1.3 Year-wise division of words collected in the TDIL corpus

Year	Words from books	Words from newspapers	Words from magazines	Words from other sources	Total words
1981	2.00	0.70	0.20	0.10	3.00
1982	2.00	0.70	0.20	0.10	3.00
1983	2.00	0.70	0.20	0.10	3.00
1984	2.00	0.70	0.20	0.10	3.00
1985	2.00	0.70	0.20	0.10	3.00
1986	2.00	0.70	0.20	0.10	3.00
1987	2.00	0.70	0.20	0.10	3.00
1988	2.00	0.70	0.20	0.10	3.00
1989	2.00	0.70	0.20	0.10	3.00
1990	2.00	0.70	0.20	0.10	3.00
Total	20.00	7.00	2.00	1.00	30.00

The table, however, hides some complexities relating to data collection that would arise from the subject-based selection of textbooks, year-based selection of newspapers, and title-based selection of magazines and periodicals.

1.6 Selection of Newspapers

If the amount of data proposed in Table 1.3 is to be collected for developing a corpus, it implies that only 70 thousand words are to be taken from a particular year taking all the newspapers published in that year into consideration of equal representation. In reality, this is a quicksand that can put a corpus designer into trouble as the following calculation shows. Let us begin with one newspaper only.

No. of pages of a newspaper	8
No. of words in each page	5000 in average (incl. advertisements)
No. of words in a newspaper in a single day	40,000 ($5000 \times 8 = 40{,}000$)
Total no. of copies of a newspaper in a year	365
Total no. of words in a newspaper in a year	14,600,000 ($40{,}000 \times 365$)

This shows that in a single year the total number of words available from a single newspaper having 8 pages is 14,600,000 (tentative). Now, if a language has five newspapers, the total number of words in a year is around 73,000,000 ($14{,}600{,}000 \times 5$) out of which we have to take only 70 thousand words. This is not an easy game for a corpus designer. The terror of statistics can tell upon their nerves, no doubt!

There exist some easy solutions to this problem, however. The selection of a single daily copy to represent the whole year is one of them. Even then, there are some challenges. Since the total number of words ($40{,}000 \times 5 = 200{,}000$) in five newspapers for a single day exceeds the total amount of words to be included in the

Table 1.4 Words collected from newspapers in a year for the TDIL corpus

Newspapers	Year	Month	Copy	Words
Newspaper 1	1990	January, February, March	1	15,000
Newspaper 2	1900	April, May, June	1	15,000
Newspaper 3	1990	July, August, September	1	15,000
Newspaper 4	1990	October, November, December	1	15,000
Others	1990	January to December	1	10,000
Total			5	70,000

corpus, we have to be highly selective in the choice of texts included in newspapers. It is rational to collect a limited number of words from each newspaper to achieve the target of 70 thousand words allotted for each year. The data given below (Table 1.4) presents an impression about how we have been able to collect data from newspapers for the TDIL corpus.

One can argue that the sampling method is not error-free since such a tiny amount of data cannot represent the uniqueness of a language use reflected in the newspapers. This is true. We also admit that we require much larger collection of text samples to understand the patterns of language use in the newspapers. However, since there was a constraint in the collection of text samples from newspapers, the proposed method proved to be the most useful strategy. In spite of many limitations, this method made two important contributions.

(a) It provided an insight to look into the problems of corpus development from a real perspective than we thought before.
(b) The gap existing between our need and actual availability of text samples provided important direction for building a corpus of newspaper texts in a more representative manner.

The selection of text samples from periodicals, journals, magazines, pamphlets, manifestos, etc., was decided by the year of publication as well as by requirement of data. Special care was taken for the language of advertisements published in newspapers, magazines, and other printed materials. Because of the uniqueness of the language, each advertisement was taken in full details and was stored in a separate text database.

1.7 Selection of Books

The selection of books published within the prescribed time period was an easier task than the troubles we had to face in the selection of texts from newspapers. Necessary help and guidance were available from the book lists published during this period which provided appropriate information about the list of published books in various subject areas and disciplines in different years. Although book lists were

available, the actual availability of books mostly depended on the support and supply of personal collections and public libraries. The personal collection was much useful in supplying books relating to music, animal husbandry, dance, cooking, knitting, sewing, beautification, and similar other subject areas besides imaginative texts like fictions, stories, and travelogues. The school and college libraries were good sources for supplying textbooks on various subjects of social and natural sciences, commerce, and other areas. The books on engineering, medical science, law, and legal activities were mostly collected from the students who were studying these subjects at the graduate and higher levels. Moreover, personal collections of books of some experts relating to specific subjects and disciplines also contributed to the task of textbooks collection.

A simple count could show that the total number of books published within a decade in a language like Bangla, Hindi, or Tamil is quite enormous. It is in the order of hundred thousand titles. Even if we could keep aside the books that were published as informative text (e.g., social science, natural science, commerce, engineering, medical science, legal texts), the number of books that were published as imaginative text (e.g., novels, fictions, stories, humors, plays, travelogues) is too large to be included in a corpus meant to contain limited number of words. That means the selection of only a few books from a huge collection is a herculean task, which needed sensible manipulation of the whole resources.

There was not much scope for the corpus designers in the act of selecting books relating to disciplines like agriculture, art and craft, social science, natural science, medicine, engineering, technology, commerce, banking, and others. Whatever books found within the range of selected disciplines and published within scheduled time period, were considered suitable for inclusion in the corpus. In certain contexts, however, some subject-specific textbooks, which were prescribed in various school and college syllabuses, were considered suitable for the corpus.

A similar method had been followed for books of history, geography, philosophy, political science, life science, physical science, commerce, culture, heritage, etc. In most cases, the method was faithful in maintaining a balance across all text types as well as in achieving the desired amount of text data for required text representation. It was noted that if this method was followed, each chapter of the books dealing with different topics marked with specific words, lexemes, terms, jargon, epithets, phrases, proverbs, idiomatic expressions, and other linguistic properties was best reflected in the corpus. The entire process of collection of text data from books may be understood from the following graphic representation (Fig. 1.2).

The application of various statistical strategies, scientific methods, and practical considerations helped the corpus designers to maintain balance, multidirectionality, and representativeness—three properties considered indispensable for the TDIL corpus which was supposed to be monolingual, comparable, general, representative, and multifunctional for the Indian languages.

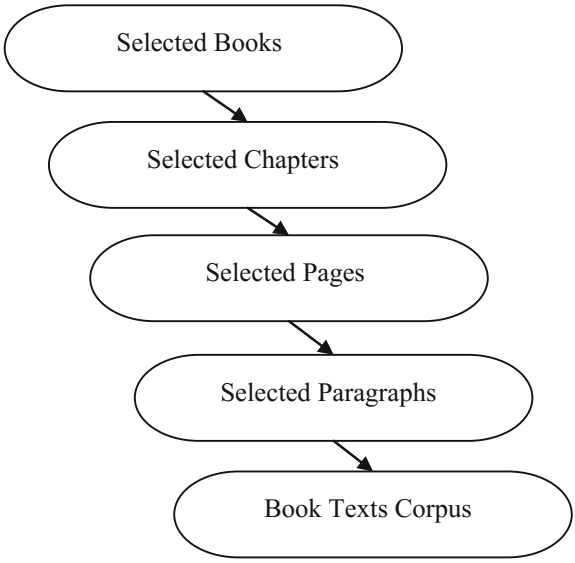

Fig. 1.2 Representation of book texts in the TDIL corpus

1.8 Selection of Writers

The selection of appropriate writers was another important issue, which arrested careful attention on the part of the corpus designers. Generally, the type of corpus we intended to develop controlled the issue of selection of text writers. That means, for instance, if we aim at focusing on the language used by woman writers, then only the texts composed by women writers are to be included in the corpus. The same approach is relevant for other corpora that are developed to represent language used in specific domains (e.g., language used by children, language in medical texts, language in legal texts, language in adult jokes). This implies that the selection of writers is a vital issue, avoidance of which may make a corpus one-sided and skewed in representation of a language (Biber 1993).

The debate that often put us in a dilemma was that whose texts should be there in the TDIL corpus? Should it contain the texts that are produced by highly acclaimed and well-known writers? Or should it contain texts produced by multitudes of less known writers? Scholars like Sinclair (1991: 37) argue that texts composed by renowned authors should hold the major share of a general corpus, since these writers, due to their popularity, larger readership, and wide acceptance, often control the pattern of use of language. Moreover, their writings are considered to be of high standard and as good representative examples of the 'right use' of a language.

On the other hand, people like us, who do not agree with this kind of approach, like to argue that the basic purpose of a general corpus is not to highlight what is acceptable, good, or right in a language, but to represent how a language is actually used by multitudes of common language users. Therefore, irrespective of any criterion of acceptance, popularity, goodness, etc., we argue that a general corpus should

include texts composed by all types of writers coming from all walks of life. Leech, a staunch supporter of this approach, argues that samples taken from a few great writers only cannot probably determine the overall general standard of a language. Therefore, we should pay attention to the texts that are produced by most of the ordinary writers, because they are not only larger in number but also more representative of the language at large (Leech 1991). We subscribed this argument and adopted a real 'democratic approach' in the selection of writers for the TDIL corpus.

1.9 Selection of Target Users

Finally, the question of target users has to be solved before the process of corpus generation starts. In many cases, predetermination of the target users can settle many of the confusions with regard to content and composition of a text corpus. In our opinion, it is necessary to identify target users due to the following reasons:

(a) The use of a corpus is not confined within natural language processing only. It has application relevance in many other fields of linguistics also.
(b) People working in different fields of human knowledge require different kinds of corpus for their specific research and application.
(c) Predetermination of target users often dissolves many issues relating to theme, content, and composition of a corpus.
(d) Target users are often relieved from the lengthy process of selection of appropriate corpus from a digital corpus archive for their works.

The form and content of a corpus may vary based on the type of corpus users. In essence, the event of corpus generation logically entails the question of possible use in various research activities and applications. Each research and application is unique of its kind that requires specific empirical data for investigation and analysis. For instance, in language teaching, a language teacher requires a learner corpus than a general corpus to suffice his/her needs. Similarly, a person working on language variation across geographical regions needs a dialect corpus than a general corpus to substantiate his research and investigation. A lexicographer and a terminologist require both a general corpus and a diachronic corpus. A speech researcher requires speech corpus. That means application-specific requirements cannot be addressed by data stored in a general corpus. Hence, the question of selection of target users becomes pertinent.

Although prior identification of target users is a prerequisite in corpus generation, it does not mean that there is no overlap among the target users with regard to utilization of a corpus. In fact, past experience shows that multifunctionality is a valuable feature of a corpus due to which it attracts multitudes of users from various fields. Nobody imagined that the *Brown Corpus* and the *Lancaster-Oslo-Bergen (LOB) Corpus*, which were developed in 1961 to study the state of English used in the USA and in the UK, respectively, would ever be utilized as highly valuable resources in many other domains of language research including English Language

Teaching (Hunston 2002) and culture studies. This establishes the fact that a corpus designed for a group of specific users may equally be useful for others. Thus, a diachronic corpus, although best suited for dictionary makers, might be equally useful for semanticists, historians, grammarians, and for people working in various branches of social science. Similarly, a corpus of media language is rich and diverse enough to cater the needs of media specialists, social scientists, historians, sociolinguists as well as language technologists.

1.10 Conclusion

At the time of generation of the TDIL corpus, the general assumption was that the proposed corpus of the Indian languages would be used in various works by one and all in linguistics and other domains. Since it is a general corpus, the number and types of users should be boundless. The main application of the corpus was visualized in the works of natural language processing and language technology. For some, it was supposed to be used in all kinds of mainstream linguistic studies (Dash and Chaudhuri 2003). In course of time, it has established its functional relevance in dictionary making, terminology database compilation, sociolinguistics, historical studies, language education, syntax analysis, lexicology, semantics, grammar writing, media text analysis, spelling studies, and other domains. As a result, over last two decades, people from all walks of life have been interested in the TDIL corpus which contains varieties of texts both in content and texture (Dash 2003).

In this present chapter, we have tried to discuss those issues which generally crop up when one tries to develop a text corpus from printed text materials for the less resourced languages. It may happen that some of the issues discussed here are also relevant for the resource-rich languages while other issues are not relevant at all. The importance of this chapter may be realized when information furnished in it becomes useful for the new generation of corpus developers who may adopt different methods and approaches based on the nature of text, nature of text source, and nature utilization of corpus data in linguistics and other domains.

References

Apresjan, J., Boguslavsky, I., Iomdin, B., Iomdin, L., Sannikov, A., and A., Sizov. 2006. A Syntactically and Semantically Tagged Corpus of Russian: State of the Art and Prospects. In *Proceedings of LREC*, 1378–1381. Genova, Italy.

Atkins, S., J. Clear, and N. Ostler. 1992. Corpus Design Criteria. *Literary and Linguistic Computing* 7 (1): 1–16.

Barnbrook, G. 1998. *Language and Computers*. Edinburgh: Edinburgh University Press.

Biber, D. 1993. Representativeness in Corpus Design. *Literary and Linguistic Computing* 8 (4): 243–257.

Cheng, W. 2011. *Exploring Corpus Linguistics: Language in Action*. London: Routledge.

Crawford, W., and E. Csomay. 2015. *Doing Corpus Linguistics*. London: Routledge.
Dash, N.S. 2001. *A Corpus-Based Computational Analysis of the Bangla Language. Unpublished Doctoral Dissertation*. Kolkata: University of Calcutta.
Dash, N.S. 2003. Corpus Linguistics in India: Present Scenario and Future Direction. *Indian Linguistics* 64 (1–2): 85–113.
Dash, N.S. 2007. Indian Scenario in Language Corpus Generation. In *Rainbow of Linguistics: Vol. 1*, ed. N.S. Dash, P. Dasgupta, and P. Sarkar, 129–162. Kolkata: T. Media Publication.
Dash, N.S., and B.B. Chaudhuri. 2003. Relevance of Corpus in Language Research and Application. *International Journal of Dravidian Linguistics.* 33 (2): 101–122.
Gries, S.T. 2016. *Quantitative Corpus Linguistics With R: A Practical Introduction*. London: Routledge.
Hunston, S. 2002. *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.
Kennedy, G. 1998. *An Introduction to Corpus Linguistics*. New York: Addison-Wesley Longman Inc.
Leech, G. 1991. The state of the art in corpus linguistics. In *English Corpus Linguistics: Studies in Honour of J. Svartvik*, ed. K. Aijmer and B. Altenberg, 8–29. London: Longman.
McEnery, A., R. Xiao, and Y. Tono. 2005. *Corpus-Based Language Studies: An Advanced Resource Book*. London: Routledge.
McEnery, T., and A. Hardie. 2011. *Corpus Linguistics: Method, Theory, and Practice*. Cambridge: Cambridge University Press.
Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
Summers, D. 1991. *Longman/Lancaster English Language Corpus: Criteria and Design*. Harlow: Longman.
Vandelanotte, L., K. Davidse, C. Gentens, and D. Kimps. 2014. *Recent Advances in Corpus Linguistics: Developing and Exploiting Corpora*. Amsterdam: Rodopi.
Weisser, M. 2015. *Practical Corpus Linguistics: An Introduction to Corpus-Based Language Analysis*. London: Wiley-Blackwell.

Web Links

http://ildc.in/Bangla/bintroduction.html.
http://tdil-dc.in/index.php?option=com_vertical&parentid=58&lang=en.
http://www.tandfonline.com/doi/pdf/10.2989/16073610309486355.
https://wmtang.org/corpus-linguistics/corpus-linguistics/.
https://www.anglistik.uni-freiburg.de/seminar/abteilungen/sprachwissenschaft/ls_mair/corpus-linguistics.
https://www.press.umich.edu/pdf/9780472033850-part1.pdf.
https://www.slideshare.net/mindependent/corpus-linguistics-an-introduction.
https://www1.essex.ac.uk/linguistics/external/clmt/w3c/corpus_ling/content/introduction.html.
https://www.wiley.com/en-us/Practical+Corpus+Linguistics%3A.
https://www.futurelearn.com/courses/corpus-linguistics.

Chapter 2
Process of Text Corpus Generation

Abstract In this chapter, we shall argue and try to explain our claims that a text language corpus, which is being developed for representing a natural language, should be large in the amount of data, multidimensional in text composition, and maximally representative to record all kinds of text varieties found in the language. If we can follow this method, a corpus will be able to overcome the pitfalls of skewedness in text representation and imbalance in text content. The achievement of these properties requires proper treatment of text on the part of the corpus designers as long as the task of data collection and storage continues. In essence, the process of text corpus generation involves various issues like methods of text data selection, hardware to be used in corpus generation, methods adopted for text entry for digitization, nature of corpus generation, the process of corpus data management, and others. We shall discuss some of these issues in some details in this chapter keeping the TDIL Indian language corpus into our purview.

Keywords Method of text selection · Technical requirements · Hardware Software · Tools · Methods of word entry · Electronic source · Internet Optical character recognition (OCR) system · Manual data entry · Corpus generation · Corpus management · Copyright

2.1 Introduction

It is already argued (Chap. 1) that a language corpus should be very large in data, multidimensional in composition, and representative in content to overcome the pitfalls of skewedness and imbalance (Barnbrook 1998). Besides, the goal for all-round representation of infinite potential varieties available within a language requires a corpus to be large and widely representative. This is an essential observation relevant to not just corpus linguistics only but to any form of scientific investigation, which relies on analysis of empirical language data (Sasaki 2003).

In a sequential order, after due consideration of various issues involved in text corpus generation, the real task of language data collection starts. This part also

N. S. Dash and L. Ramamoorthy, *Utility and Application of Language Corpora*,
https://doi.org/10.1007/978-981-13-1801-6_2

involves various other issues that require proper treatment on the part of the corpus designers as long as the task of data collection, storage, and distribution of data continues. In general, it involves issues like the method of text selection, hardware requirements, methods of text entry, electronic source, manual data entry, corpus generation, corpus management, copyright. We shall also discuss these issues with reference to the TDIL Indian language corpus.

In Sect. 2.2, we shall discuss the methods of text selection for a corpus; in Sect. 2.3, we shall address the technical requirements for corpus generation; in Sect. 2.4, we shall describe different methods of word entered into a corpus; in Sect. 2.5, we shall highlight the process of actual text corpus generation; in Sect. 2.6, we shall address the issues and challenges involved in corpus data management; in Sect. 2.7, we shall try to know what kind of language data and linguistic information a corpus can supply to us; and in Sect. 2.8, we shall refer to some of the major issues relating to the problems of copyright of texts in corpus data in general.

2.2 Method of Text Selection

There are some formality indices for the selection of samples from various text categories to be included in a written text corpus (Sigley 1997). For developing the TDIL corpus for Indian languages, however, we have depended on the texts of regularly used Indian languages based on relative frequency of readership of various printed texts by common Indian population. From a pilot sample survey conducted for this purpose, we have found that the common Indian people within the range of 30–60 years of age usually read printed texts in the following order (Table 2.1).

Table 2.1 gave us an interesting insight into the reading habit of the Indian population in general. It showed that the Indian people most often tend to read newspapers

Table 2.1 Distribution of text-based readership of the Indian population

No.	Text categories	Percentage (%)
1	Newspapers published in vernaculars	40
2	Newspapers published in English	05
3	Books and papers on arts and humanities	10
4	Books and journals on social sciences	10
5	Books and journals on natural sciences	07
6	Books and journals on medical sciences	05
7	Books and journals on engineering and technology	04
9	Books and journals on business and commerce	10
10	Books and journals on legal and administration	03
11	Other texts (periodicals, magazines, manifestos, etc.)	06
	Total	100

Table 2.2 Amount and percentage of text from each text category in the TDIL corpus

No.	Text categories	Words	Percentage (%)
1	Creative writing texts	300,000	10
2	Fine arts texts	300,000	10
3	Social science texts	300,000	10
4	Natural science texts	300,000	10
5	Medical science texts	300,000	10
6	Technology and engineering texts	300,000	10
7	Mass media texts	300,000	10
8	Commerce and industry texts	300,000	10
9	Legal and administration texts	300,000	10
10	Other texts	300,000	10
	Total	3,000,000	100

published in vernaculars than any other printed text materials. In fact, a great majority of Indian population read only newspaper texts and nothing else, while many others read only the 'transient texts' found in periodicals and magazines. Many people, if they are not professionals or engaged in some sort of academics, do not read scientific, technical, medical, and administrative texts. The table also showed that the percentage of people reading books and newspapers belonging to arts and humanities, social sciences, and business and commerce are of the same range, while the percentage of people belonging to the readership of other kinds of text is very low. What was most important for us was that people liked to read newspaper texts the most followed by imaginative texts like fictions and novels. They had shown least preference for the books and other texts materials of social sciences, general knowledge, natural sciences, business and commerce, and translation.

The opinions of students studying in schools and colleges are not taken into consideration since the members of this group usually read those books and texts, which are prescribed to them to read. In most cases, reading habit of this group of people is normally controlled not by their personal choice but by a motivation of passing examinations successfully. Therefore, any consideration of reading habit of this group would have definitely affected the result of the survey. Moreover, since the basic aim of the survey was to know about the kinds of printed materials the common people usually include in their reading habits, the result of reading habits of the students was not of much relevance here.

The results of this survey helped the corpus designers to classify all available printed text materials, according to the nature of texts, into nine major text categories with a goal for collecting a fixed number of words from each category for the TDIL corpus (Table 2.2).

Each text category is further divided into several subcategories according to the type of the text included in it, in the following manner (Table 2.3).

Table 2.3 Text categories and their subcategories in the TDIL corpus

Broad categories	Subcategories
Creative writing	Fictions, novels, short stories, essays, travelogues, and others
Fine arts	Painting, drawings, music, films, sculpture, and others
Social science	History, political science, sociology, linguistics, etc.
Natural science	Physics, chemistry, biology, zoology, botany, etc.
Medical science	Allopath, homeopathy, ayurveda, naturopathy, etc.
Technology and engineering	Computer, electronics, mechanical engineering, etc.
Mass media	Newspaper, advertisement, magazines, etc.
Commerce and industry	Banking, finance, housing, management, etc.
Legal and administration	Legal notices, public circulars, government orders,
Others	Translation, manifesto, placards, festoons, etc.

In this context, it should be mentioned that text samples from poetry, verse, songs, rhymes, riddles, ballads, and limericks and other poetic sources were not included in the TDIL written corpus of prose texts. The reasons behind keeping these text samples separate from the prose texts are as follows (Dash 2006).

(a) The structure and treatment of language used in poetic texts are different from the language used in prose texts (e.g., scientific texts, newspaper texts).
(b) The texts used in poetry vary from that of prose not only in lexical choice and word use but also in word formation, sentence formation, lexical collocation, use of multiword units, etc.
(c) Quite often, the texts taken from poetry show that the terminal verb is used either at the beginning or at the middle of a sentence. This is a quite irregular pattern of use of the verb in most of the poetic texts in the Indic languages.
(d) In case of a language like Bangla, both *sādhu* (i.e., chaste) and *calit* (i.e., colloquial) forms of words are quite randomly intermixed in poetic texts. This kind of use of words is hardly noted in case of prose texts.
(e) In case of poetic texts, it is also noted that sometimes the original structures of words are changed to maintain metrical balance in rhymes, which has never happened in case of prose texts.

These and many other characteristic features of the language of poetry have motivated the corpus designers to keep texts of poetry away from the corpus of prose texts. This is a common phenomenon in all the languages of the world. In fact, the presence of certain unique linguistic phenomena in poetic texts inspires us to argue for generating 'poetry corpus' as a separate type of corpus. A poetry corpus can be analyzed separately to understand its own unique linguistic forms and features (Coleman and Kay 2002). We also argue that we should seriously try to develop both synchronic and diachronic corpora of poetic texts so that these can be used with prose text corpora (as comparable corpora) to mark out the notable traits of differences between the prose text and poetic text used in a language across centuries.

2.3 Technical Requirements

The creation of a text corpus in the Indian languages requires a special kind of technical support with innovative software and system. For this task, the corpus designers require some unique devices and tools to deal with Indian scripts and their complexities because most of the Indian languages scripts (barring a few national languages) are not yet Unicode compatible and Internet-friendly. As a result of this, texts of these languages are not available in digital form (Dash 2003). At the early years of corpus development in the Indian languages, the corpus developers used the following technical devices for their purposes:

(a) Several personal computers with good storage capacity,
(b) Transcript Card based on Graphics and Intelligence-based Script Technology (GIST),
(c) An ISCII to ASCII converter,
(d) A software called Script Processor (SP),
(e) Several display monitors,
(f) Conventional computer keyboards tagged with Indian scripts,
(g) Multilingual printers,
(h) Floppy diskettes.

In 1985, the *Indian Institute of Technology* Kanpur developed a multilingual computer terminal technology, which was later modified at the *Centre for Development of Advanced Computing* (C-DAC), Pune, resulting in the development of the Graphics and Intelligence-based Script Technology (GIST) terminals for the Indian language scripts. These tools and technology were used for displaying the scripts of various Indian languages on the computer screen based on the information entered through standard English keyboard. The interface carried an overlay of characters used in respective Indian scripts for display and ease in data entry. Based on these technology developments, the codes for various keys were accessed for the Indian scripts and their layouts had been standardized by the Bureau of Indian Standards (BIS), Government of India (Murthy and Despande 1998: 7). Thus, the text corpora in the Indian languages were developed in those days with the help of a Transcript Card installed within a personal computer.

The technology incorporated within the Transcript Card was capable to display and print the texts in the Indian scripts as and when required by the corpus designers. The Transcript Card is a hardware add-on card that has been used to update the IBM PC/XT/AT compatible for graphical interactions in all major Indian scripts, besides English. With the installation of the Transcript Card inside a personal computer, the corpus designers could use almost the entire range of standardized English interpreters, text compilers, and text-oriented application packages. With this interface, the corpus designers were also able to input, retrieve, and print texts in any of the Indian language scripts they required. Moreover, following the American Standard Code for Information Interchange (ASCII), the Transcript Card was used to design the Indian Standard Code for Information Interchange (ISCII) which was later rec-

ommended and adopted by the Bureau of Indian Standards (BIS), Government of India (ISI Code No. IS 13194: 1991).

The Script Processing software built into the Transcript Card provided useful technical support for processing texts in all the Indian scripts in a uniform manner. Moreover, it provided a simple user interface as well as facilitated a possibility for combining both the Indian and the Roman (English) scripts within a single text document. The existing standard keyboards of the IBM compatible at PCs/XTs/ATs were used with Transcript Card after putting labels for letters used in the Indian scripts on the top of the keys assigned to English letters.

The interface provided the two display facilities on the monitor of a computer: One was displayed in the conventional English character mode, and the other was displayed in Indian multilingual mode. It used 8-bit ISCII code, which also contained a 7-bit ASCII code in its lower half for display. The upper 128 character positions in the keyboard were used for the characters used in each Indian script, while the lower 128 character positions are used for the English (Roman) script. Thus, the Transcript Card had enabled the corpus designers to store language texts written in the Indian scripts in computer and use these data in various works according to their needs.

Although the achievement was considered as one of the important milestones in the field of language technology in the Indian context, the Transcript Card had exhibited some of its limitations, which created problems both in text data input and text processing. Therefore, it was essential to identify the unique features as well as limitations of the software that may help the system developers to enhance its functional efficiency, operational robustness, and application potentials (Dash 2005: 151). However, due to recurrent use of the software on a large scale across the country over a decade, some modifications were possible to incorporate in the new version of the interface. As a result, those who started working with the new version of the software did not face much trouble with it.

Thanks to the development of script technology within last few years, the prose texts in many of the Indian language scripts are now available on the Internet. At present, if anyone is interested in developing a text corpus in any of the Indian languages, it will not be an uphill task for him. However, one has to verify in which format the texts are available on the Internet. The present state of availability of texts in the Indian language scripts is in two formats:

(a) Texts available in .pdf format and
(b) Texts available in the Unicode format.

Majority of texts in the Indian scripts are available in .pdf format. This is because, for most of the Indian language scripts, particularly of those languages which are not included in the 8th Schedule of the Constitution of India, there is no Unicode. Therefore, the preparation of texts in digital form in these languages is a real tough task. What people usually do in these situations is that they type texts in a character layout developed arbitrarily, convert these texts into .pdf files, and put these on the Internet. The development of a digital corpus with texts from these materials is nearly impossible as the conversion of .pdf files into Rich Text Format (RTF) or .doc files generates nothing but pure garbage data which is useless in any research and

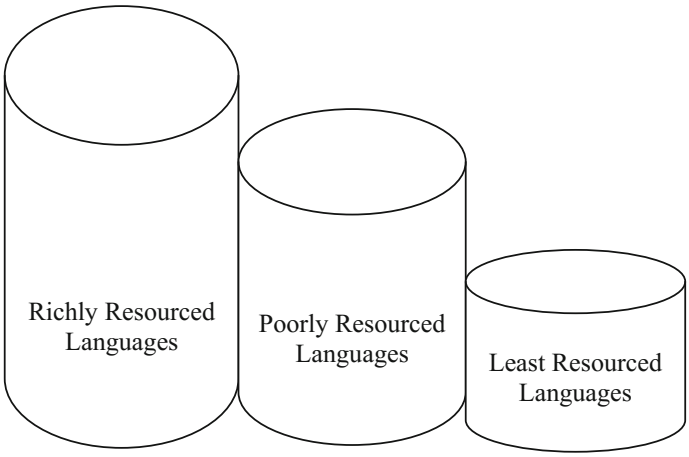

Fig. 2.1 Division of languages based on digital text resources

development works. Until good script technology and Unicode are developed for these languages, the development of language corpus, digital resource, and language technology for these languages will remain an elusive dream.

The other kinds of texts that are available on the Internet are in Unicode format. These are mostly from those languages which are included in the Constitution of India and which have already established their linguistic majority in the country. Languages like Hindi, Bangla, Punjabi, Marathi, Urdu, Tamil, Telugu, Malayalam, Kannada, Odia, Assamese, Kashmiri, Gujarati, Konkani, and others have made their existence visible in the Internet (although in limited scale) with varied amount of text data in Unicode as well as in .pdf versions. Collection of text data from these languages from the Internet is comparatively simplified, in the sense that if someone tries, he or she will not end up with empty hand, rather he or she will be able to collect at least some amount of text data in Unicode version, even though it is known to all of us that all the major Indian languages, from digital perspective, are yet considered as 'resource-poor' languages when these are compared with English, French, Chinese, German, Spanish, and Japanese which are considered as 'resource-rich' languages (Fig. 2.1).

We need to have a new set of technological devices if we are interested to collect text data from these languages from the Internet and other digital sources. We have envisaged this because the methodology and the technology that we adopted and used for the generation TDIL text corpus nearly 30 years ago are no more usable now. There have been several changes over the years in the form of technology and insights into the act of text corpus generation and processing due to which any present act of corpus collection must keep in mind the nature of paradigm shift reflected in the following orders (Table 2.4).

Table 2.4 Nature of paradigm shift in corpus generation in India

No.	During 1991	At present (2018)
1	Availability of printed texts	Availability of printed texts
2	Non-availability of digital text	Availability of digital texts
3	No Unicode-compatible Indian scripts	Unicode-compatible Indian scripts
3	ISCII (ISCII)-based Indian texts	Unicode-based Indian texts
4	No homepage/Web sites with Indian language text materials	Many homepage/Web sites with Indian language text materials
5	GIST Technology	HTML Technology
6	Less diversity in subject materials	The wide diversity in subject materials
7	No language data collection tool or technology	Many language data collection tool or technology
8	No NLP tools and technology	Many NLP tools and technology
9	Zero public awareness	Much public awareness
10	No experience in corpus generation	Long experience in corpus generation
11	No trained NLP scientists	A good number of trained NLP scientists
12	No vision about the application of language data	Clear vision about the application of language data
13	Marginal or sparse government or private support	Marginal or sparse government or private support

Now, the situation is that if someone wants to develop a corpus both from printed and digital sources in any of the Indian languages that have a digital presence on the Internet, one may require the following tools and devices:

(a) Language data access devices,
(b) Language data collection devices,
(c) Language data storage devices,
(d) Language data management devices,
(e) Language data processing devices,
(f) Language data analysis devices,
(g) Language data distribution devices,
(h) Language data utilization devices.

For all these activities, we require appropriate hardware devices (e.g., computers, desktops, laptops, printer, scanner, CDs, pen drives, external hard disks) as well as appropriate software (e.g., NLP toolkits, SPSS, Text Processing Tools). Moreover, we need a large team of trained linguistics (more specifically, computational linguists) who have good command both in linguistics and computation.

2.4 Methods of Word Entry

It is clear that the technique that we can use for text collection from a less resourced language will vary from the techniques that we use for richly resourced language. Similarly, the processes that can be used for text data collection may vary from that of a speech data collection, since each corpus type is characteristically different from the other. Moreover, due to basic differences underlying between two text types, a collection of words will also vary in two different ways. While samples of speech data may be acquired from the speech of the informants directly with an aid of tape recorder and similar such electronic data recording devices (Sinclair 1991: 14–15), samples of written text data can be collected from printed and digital text documents and materials (Kennedy 1998: 78–80) following the five methods classified below:

(1) Data from electronic sources,
(2) Data from World Wide Web(WWW),
(3) Data from emails and tweets,
(4) Data from machine reading of texts, and
(5) Data from manual text entry.

2.4.1 Data from Electronic Sources

It is possible to gather text samples from various types of electronic sources to generate a text corpus. That means it possible to include texts produced and published by newspapers, magazines, books, periodicals, journals, etc., if these texts materials are made available in digital version. At present, several Indian languages have made these materials available on the Internet either in HTML text format or in .pdf version. People can collect a large amount of data from these digital sources to develop a good-quality text corpus based on their requirement within a short span of time. They, however, need a good .pdf converter so that all the .pdf text materials are rendered into a .doc version for further processing, analysis, and application.

In case of Unicode supported languages, a major amount of data of the *British National Corpus*, the *American National Corpus*, the *International Corpus of English*, etc., have been collected by way of gathering data from electronic sources. Following the same method, corpora in many other less resourced languages are generated, which are large in a number of words, diverse in text types, and useful for object-oriented and general research and applications. However, this method is not fully applicable for many of the Indian languages, since there are very limited amounts of text materials available in electronic form. The scenario, however, is changing rapidly, and we expect that within next few years most of the Indian languages will have a large amount of texts in electronic form. At present, whatever text materials available in electronic form can be collected to compile a corpus of various types and applications.

2.4.2 Data from WWW

For advanced languages, an enormous amount of text data is available in a digital version from various Web pages, Web sites, home pages, etc. A modern corpus developer can easily collect these text materials by using a good and customized crawler system to gather data for corpus generation. In fact, it has been reported in many research works how the researchers have collected text data for English, German, French, Chinese, Japanese, and Spanish from various Web sources by using a crawler. Moreover, they are using these data as a corpus for language technology works.

For less resourced languages, as it happens for most of the Indian languages, this may not be a useful process as the amount of data in WWW sources of these languages is not large and varied. Moreover, the crawling tools and devices that work wonders for many of the advanced languages are not so effective for most of the Indian languages (Dash and Arulmozi 2017: 126). Therefore, a collection of text data from these sources for the Indian languages is a challenging task that can be achieved only when a large amount of text data are made available in the cyber world and scientists develop good crawling devices for the Indian languages for collecting data from these sites.

2.4.3 Data from Email and Tweets

Countless text materials in the Indian languages are available from digital gateways like emails, tweets, blogs, and similar other sources. Although texts produced in these digital mediums are not of the same nature and kind as we find them in serious academic and scientific materials, one can still like to use these text materials to generate a corpus of a different type in Indian languages. In fact, it is possible to generate a large multidimensional corpus of various text types within a short period of time using a limited amount of resources and manpower if one collects data from these sources. Although this is not a powerful method for text corpus generation, it has been of much use for many of the advanced languages and can be equally useful for many Indian languages. Today, there are good facilities by which we can send emails and Web mails or tweet in many Indian language scripts. The software giants have made path-breaking success in this area for us, and we can use the technology to serve our goals for corpus generation in the Indian languages.

2.4.4 Data from Machine Reading of Texts

Automatic machine reading for printed texts and their subsequent conversion into electronic text with the help of an optical character recognition (OCR) system is a

useful method for text corpus generation. The materialization of this process, however, requires a system by which we can convert printed pages into machine-readable images which can be converted into electronic texts by the OCR software. For the advanced languages, at the early stage, an OCR was a regular method of corpus generation. At present, they do not require it much because the amount of text data is so vast and varied on the Internet that they hardly require an OCR system for generating a text corpus. They use this software only for converting old printed and handwritten texts into a digital version.

For the Indic languages, it is possible that by applying an OCR technology a large number of printed materials can quickly be converted into electronic form. In that case, we need a robust OCR technology that can effectively serve our purposes. The reality is, however, not so promising as good-quality OCR software is not yet ready for many of the Indian languages, although it is claimed to be available for a few Indian languages with limited success (Pal and Chaudhuri 1998, 2004; Vikas et al. 2003; Ghosh et al. 2010; Govindaraju and Setlur 2010; Mathew et al. 2016). Whatever tools are available for the Indic languages (including Google OCR), these are not much use for the Indian language scripts. They fail to read the Indian language script properly and, therefore, generate a large number of orthographic errors in output—particularly in case of old printed text documents. Moreover, they fail miserably if the text document contains font variation, mutilated materials, handwritten texts, and multiscript texts.

2.4.5 Data from Manual Entry

The process by which the largest amount of written texts data may be converted into a digital corpus for the Indian languages is simple manual data entry in a computer with the help of an Indic language interface. So far, this has been the most useful way of corpus generation from many of the Indian languages. By using this method, people have succeeded collecting text from sources like printed books and papers, handwritten texts materials, transcripted spoken texts, personal letters, old manuscripts, handwritten diaries, old bond and wills, and many other text sources.

For this task, a few expert data entry operators are employed who can type in words into the computer from the printed text materials available to them. Although this is a very old-fashioned system, which consumes much time, money, and energy, it is a far better process and highly effective for those Indic languages where electronic texts are not available. At the time of TDIL Indic text corpus generation, due to non-availability of other techniques, the Indian corpus designers had to follow this method to build corpora in the Indian languages (Dash 2007). Even today, many of the people use this method for generating text corpus in many of the Indian languages including both standard and local varieties.

In this context, it may be mentioned that some industrial agencies, commercial houses, and research centers (e.g., Linguistic Data Consortium, USA; Oxford Text Archive, UK; Lancaster University, UK; ICAME, Norway) have already developed

or started developing electronic text corpora of different types in some of the Indian languages. As required, one can procure these corpora free of cost or with payment for linguistic research and application.

2.5 The Process of Corpus Generation

In case of manual corpus collection, the actual work of word entry starts when software is activated, corpus generation interface is invoked, and a text file name is created. For technical constraints, the name of a text file is usually limited to eight characters. For the convenience of work, characters of each text file name are arranged in the following order:

(a) First two characters represent a text category.
(b) Next four characters represent the name of a text.
(c) Last two characters represent the serial number of a text file.

Immediately after the eight characters, a dot (.) is placed followed by a three-letter abbreviated form representing the name of an Indian language. For instance, consider the file name: NLKLBL05.BAN, where 'NL' stands for text category (Novel = NL), 'KLBL' stands for the title of the text [i.e., *kālbelā* (name of a novel published in Bangla)], and '05' indicates that the file is fifth in serial number created from this text. The extension 'BAN' after the dot stands for the language 'Bangla.' Each text file, created in this manner, has two major parts, as the following:

(a) **Header Part**: This is the metadata part that contains various extralinguistic information of a text (e.g., name of a book, year of publication, edition, name of the author(s), name of the publisher, number of pages taken for input, type of a text) that is required for maintaining text records, managing text data, classification of texts, disseminating corpus data, dissolving copyright problems as well as for reference to sociolinguistic and stylistic works.
(b) **Text part**: This part contains original text in the Indian languages. Manual input of physical text begins from the second line in the Indian scripts following the Unicode installed in a computer and a keyboard.

At the time of manual input, the physical width of the lines of printed texts is normally preserved in the digital version of the text. Generally, a physical line of a printed text spreads within a range of 80–100 characters, while the actual screen line of a monitor can extend more than 200 characters. Therefore, technically, there is no problem in keeping the width of a line of physical text intact in the digital version of a text.

After a paragraph of text thus generated, one line should be left blank to begin a new paragraph. This helps in automatic identification of paragraphs as well as for counting the number of words in a paragraph. Since text samples in the Indian languages have been collected and compiled in more or less in a random sampling

<?xml version="1.0" ?>			
<?xml-stylesheet type="text/css" href="home.css"?>			
<Doc id="BAN-W-Media-	B0035	"	Lang="Bangla">
<Header type="text">			
<encodingDesc>			
<projectDesc>	TDIL-Bangla Corpus, Monolingual Written Text		</projectDesc>
<samplingDesc>	Simple written text. Pages: 11,12, 15, 17, 24, 31, 32, 40, 41, 48, 49, 56, 64, 65, 74, 75, 79, 80		</samplingDesc>
</encodingDesc>			
<sourceDesc> <biblStruct> <source>			
</body></text> </Doc>			

Fig. 2.2 Information captured in machine-readable form in metadata

manner, a unique symbol is necessary to be used at the beginning of a new text sample to determine its separate textual identity.

For the purpose of text processing, each text file is kept in a simple format: a series of words marked with blank space and punctuation marks. While legal and copyright information, page numbers, and line numbers are preserved for reference purposes within the metadata, other information (e.g., page layout, setting, typeface, font type) are mostly ignored in the digital version of the text.

With all these information clubbed into a machine-readable format (Fig. 2.2), a computer is able to understand and identify a text that is made with a long succession of nondescript sets of characters marked off in pages and lines. Thus, following this method across all the Indian languages, more than 1200 text files are generated in the TDIL project each one of which contains around 3000 words of running texts.

2.6 Corpus Management

The management of a text corpus in digital platforms is a highly complicated task. We have noted that there are always some typographic errors to be corrected, some modifications to be in the metadata generated for a file, and some improvements to be incorporated into the technology used (Summers 1991). At the initial stage of corpus creation, there are also some other issues that are linked with the systematic arrangement of the generated text files based on various types of texts. Through the application of these processes, the subsequent works of searching language data and information will be much faster and easier.

Generally, the utility of corpus data is largely enhanced through the application of intelligent management of text files in a digital archive. It is understood that the task of retrieval of data and information from a corpus requires utmost attention on the part of the corpus users so that they can easily retrieve the required text files and text data for the oriented application. Moreover, it is understood that systematic management of text files can make interdisciplinary research more effective and less distractive. The major activities relating to the management of corpus data are the following:

(1) Corpus data storage,
(2) Metadata annotation,
(3) Header file preparation,
(4) Retrieval of text data,
(5) Up-gradation of data,
(6) Cataloguing of data,
(7) Dissemination of data.

Once a text corpus is created and stored in the system, the corpus designers need to apply necessary methods and schemes for the purpose of maintenance of texts, augmentation of data, and up-gradation of the systems. First, with regard to maintenance, the corpus designers need to keep a clear watch on the corpus so that the data is not corrupted by virus infection or physically damaged due to some external physical factors. Second, the primary process of corpus augmentation has to be continued for generations to enlarge and update the existing amount of data with new text samples that may be collected from different new sources that are being produced in a language over the years. Finally, the corpus developers have to keep an eye on the availability of new technology tools and devices for up-gradation of the corpus. It is needed because the existing text data has to be converted properly to make it usable in new systems and techniques. Since the scenario of computer technology is changing very rapidly with time, data stored in corpus has to be continuously upgraded so that these are at par with the new systems and technology available. Else, the entire corpus database will be useless, as it has happened for the TDIL corpus developed in 1995. In general, the process of up-gradation of the database in a corpus involves the following four major activities:

(a) Storage and preservation of text data in the hard disk as well as at the central data store (maybe, at the digital corpus archive) in multiple copies and versions,
(b) Transportation of text data from the hard disk to floppy disk and similar other devices (e.g., compact disks, pen drives, external hard disk, cloud source, Google drive, or similar other reliable storing devices),
(c) Making the corpus usable for all operating systems (OS) like Disk Operating System (DOS), Windows, Linux, or other operating systems, and
(d) Conversion of texts from the ISCII to the ASCII. And in the similar fashion, texts generated in ASCII have to be converted into Unicode and, if required, from the Unicode to other advanced technology of character rendering system.

In essence, the adaptation and assimilation of new hardware and software technologies are indispensable in the activities relating to text corpus generation, text processing, text analysis, and text utilization. In this case, the corpus developers have to take care of several issues, as stated, with optimum attention. Although the present state of computer and language technology is advanced and improved enough to carry out all these works with great satisfaction, it will not be a daydream to expect that within next few years the status of software technology will improve to a large extent to address many other complex issues and challenges relating to corpus generation, storage, management, and utilization.

2.7 What Does a Corpus Supply?

Once a corpus is made ready for general use, the question that is raised quite often is this: What does a corpus supply to the language users? To respond to this question, it is noted that a well-formed and well-designed general text corpus can supply the following things to the language users:

(1) Language data,
(2) Information,
(3) Examples, and
(4) Insights.

With regard to **language data**, a text corpus supplies data in a lot with various types. At the character level, it supplies data relating to letters, graphemes, allographs, diacritics, conjuncts, numerals, punctuations, and other textual symbols. At the morpheme and word level, it supplies data about single words, morphs, compound words, multiword units, word-formative elements, and named entities. At the larger structure level, it provides data relating to idioms, phrases, clauses, sentences, set expressions, and proverbs.

With regard to **information**, a text corpus supplies information of two types: (a) intralinguistic information and (b) extralinguistic information. In case of intralinguistic information, it supplies textual, intertextual, lexical, lexicological, lexicographic, semantic, syntactic, grammatical, discoursal, anaphoric, figurative, and prosodic information. In case of extralinguistic information, it provides cultural, demographic, social, historical, stylistic, deictic, ecological, ethnographic, and information of the external world at large.

With regard to **examples**, it supplies examples of all kinds including that of intralinguistic, extralinguistic, language elements, structure, construction, content, treatment, properties, usage, functions, patterns, formation, citation, reference, relations, stylistics, etc., of all the intralinguistic and extralinguistic properties included in the corpus.

Finally, with regard to **insights**, both from linguistic and other angles, a text corpus is a real eye-opener for understanding a language from which it is made. It gives deep insight into life, language, society, culture, time, place, event, agent, people, situation,

context, motive, goal, cognition, usage, competence, performance, communication, negotiation, mediation, disposition, and community. Looking through the corpus what we ultimately realize is that a language is a living ethnobiological entity the analysis of which gives us a complete picture of its speakers and its community on the axis of time and space.

2.8 The Issue of Copyright?

The generation of a text corpus eventually comes to the end with addressing the last line of conflict: Whose corpus is this? More accurately, whose data is this? This raises the question of copyright as several stakeholders appear in the scene to claim credit from the generated resource. In our view, there are five different types of people who can fight for copyright of a corpus:

(1) Authors of texts,
(2) Publisher of texts,
(3) Corpus developers,
(4) Project financers, and
(5) Corpus users.

The authors argue that they are the main producers of the texts. Therefore, all credits, profits (financial or otherwise), and copyright should go to them. The publishers of printed texts, on the other hand, argue that the credit of authors as text producers can easily be nullified on the ground that without their full technical and financial help no author can make a mark in the market. Therefore, copyright should belong to the publishers, and not to the authors, who can at best have intellectual property rights.

The corpus developers also sometimes ask for copyright, because they argue that the corpus they have developed is neither a text produced by an author nor a material published by a publisher. It is a text of a different kind with its own unique form, structure, and application. Therefore, copyright should be theirs. The financer (public, private, or government) of the project is also on the floor to claim that the entire project of corpus generation has been possible because of the fund provided by them. They, therefore, have the actual copyright on the corpus.

Finally, there are the corpus users. They are perhaps the most legitimate claimants of the copyrights. Their argument is simple: The value of a corpus is realized only when it is used. If they do not use a corpus, a corpus however fantastic it may look has zero academic, commercial, as well as referential relevance. Beauty lies in the eyes of the beholder! Therefore, the corpus users are the most pertinent copyright holders, not others.

In essence, the question of copyright is not yet dissolved. It is still a lingering issue for the much advanced languages as well as for the less advanced languages. All are in the same boat in a turbulent sea with expectation for a safe shore!

2.9 Conclusion

A simple format of a text corpus has the ability to reflect on the 'state of the art' of a 'language in use' for research, analysis, and application (Hunston 2002). Therefore, it is pertinent to keep the text data in a corpus in two ways: one version of the text in a simple format and the other version of the text with the addition of metadata information. Since each particular investigation is destined to look for different kinds of language data according to its own priority, it is better to have, at least, two versions of a corpus to cater the requirement of all possible users. This will enable both linguists and language technologists to look at the entire corpus (or a part of it) and retrieve data based on their requirements.

Within a general scheme of work, the corpus developers can think of applying corpus data broadly on the following domains.

(1) Language processing tools and techniques development,
(2) Language technology systems and devices development,
(3) Digital linguistic resources development,
(4) Translation support systems development,
(5) Man–machine interface development,
(6) Speech technology development,
(7) Mainstream linguistic analysis and description,
(8) Applied areas like lexicography and language teaching,
(9) Sister disciplines of social and humanistic sciences,
(10) E-governance and public support systems.

A highly imaginative corpus designer can probably speculate about people who can be the possible users of a particular type of corpus. But he can never be confirmed about its users because a new kind of research which is never visualized by a corpus designer may require the corpus data to be formatted from a different perspective. In that case, it is sensible that a text corpus is available in its raw and undistorted version as well as in its annotated form. Moreover, since a general corpus usually possesses the up-to-date text samples collected from the current usage of a language, information, which a subject-specific annotated corpus cannot furnish, may be easily obtained from a general raw corpus. Therefore, we like to argue that a text corpus should be made available in at least two versions (raw and annotated) keeping in mind the possibility of diverse use of corpus texts in different domains of linguistics and sister disciplines.

References

Barnbrook, G. 1998. *Language and Computers*. Edinburgh: Edinburgh University Press.
Coleman, J., and C.J. Kay (eds.). 2002. *Lexicology, Semantics and Lexicography: Current Issues in Linguistic Theory*. Select Papers from 4th G.L. Brook Symposium. Amsterdam: John Benjamins.

Dash, N.S. 2003. Corpus Linguistics in India: Present Scenario and Future Direction. *Indian Linguistics* 64: 85–113.
Dash, N.S. 2005. *Corpus Linguistics and Language Technology: With Reference to Indian Languages*. New Delhi: Mittal Publications.
Dash, N.S. 2006. Speech Corpora vs. Text Corpora: The Need for Separate Development. *Indian Linguistics* 67: 65–82.
Dash, N.S. 2007. Indian Scenario in Language Corpus Generation. In *Rainbow of Linguistics: Vol-I*, ed. N.S. Dash, P. Dasgupta, and P. Sarkar, 129–162. Kolkata: T. Media Publication.
Dash, N.S., and S. Arulmozi (eds.). 2017. *History, Features, and Typology of Language Corpora*. Singapore: Springer Nature.
Ghosh, D., T. Dube, and A.P. Shivaprasad. 2010. Script recognition—a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* 32 (12): 2142–2161.
Govindaraju, V., and S.R. Setlur (eds.). 2010. *Guide to OCR for Indic Scripts: Document Recognition and Retrieval*. London: Springer.
Hunston, S. 2002. *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.
Kennedy, G. 1998. *An Introduction to Corpus Linguistics*. New York: Addison-Wesley Longman Inc.
Mathew, M., A.K. Singh, and C.V. Jawahar. 2016. Multilingual OCR for Indic Scripts. In *Proceedings of 12th IAPR Workshop on Document Analysis Systems (DAS-2016)*, Santorini, Greece, 11–14 April 2016, 186–191.
Murthy, B.K., and W.R. Despande. 1998. Language Technology in India: Past, Present, and the Future. In *Proceedings of the SAARC Conference on Extending the Use of Multilingual and Multimedia Information Technology* (EMMIT 1998), Pune, India.
Pal, U., and B.B. Chaudhuri. 1998. A Complete Printed Bangla OCR System. *Pattern Recognition* 31: 531–549.
Pal, U., and B.B. Chaudhuri. 2004. Indian Script Character Recognition: A Survey. *Pattern Recognition* 37: 1887–1899.
Sasaki, M. 2003. The Writing System of an Artificial Language: For Efficient Orthographic Processing. *Journal of Universal Language* 4 (1): 91–112.
Sigley, R. 1997. Text Categories and Where You Can Stick Them: A Crude Formality Index. *International Journal of Corpus Linguistics* 2 (2): 1–39.
Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
Summers, D. 1991. *Longman/Lancaster English Language Corpus: Criteria and Design*. Harlow: Longman.
Vikas, O., P.K. Chaturvedi, P. Chopra, V.K. Sharma, M. Jain, and S. Chandra (eds.). 2003. *Vishwabharat: Indian Technology Newsletter 10*, July 2003.

Web Links

http://ildc.in/Bangla/bintroduction.html.
http://ota.ox.ac.uk/.
http://tdil-dc.in/index.php?option=com_vertical&parentid=58&lang=en.
http://www.tandfonline.com/doi/pdf/10.2989/16073610309486355.
https://wmtang.org/corpus-linguistics/corpus-linguistics/.
https://www.anglistik.uni-freiburg.de/seminar/abteilungen/sprachwissenschaft/ls_mair/corpus-linguistics.
https://www.futurelearn.com/courses/corpus-linguistics.
https://www.ldc.upenn.edu/.
https://www.press.umich.edu/pdf/9780472033850-part1.pdf.
https://www.slideshare.net/mindependent/corpus-linguistics-an-introduction.
https://www.wiley.com/en-us/Practical+Corpus+Linguistics%3A.
https://www1.essex.ac.uk/linguistics/external/clmt/w3c/corpus_ling/content/introduction.html.

Chapter 3
Corpus Editing and Text Normalization

Abstract In this chapter, we propose for applying processes like pre-editing and text standardization as some of the essential components of corpus editing and text normalization for making a text corpus ready for access across various domains of linguistics and language technology. Here, we identify some of the basic pre-editing and text standardization tasks, and we describe these works with reference to Bangla text corpus. As the name suggests, text normalization involves diverse tasks of text adjustment and standardization to improve utility of the texts stored in a corpus in manual- and machine-based applications. The methods and the strategies that we propose here to overcome the problems of text normalization are largely tilted toward written text corpus since text normalization activities relating to spoken text corpus usually invoke a new set of operations that hardly match with the normalization processes normally applied on written text corpus. The normalized version of a text not only reduces workload in subsequent utilization of a corpus but also enhances its accessibility by man and machine across all domains where language corpus has application and referential relevance.

Keywords Corpus · Pre-editing · Standardization · Normalization
Global readiness · Overlap · Term consistency · Metadata · Transliteration
Tokenization · Disambiguation · Frozen terms

3.1 Introduction

Pre-editing is a process of adjusting texts in a text corpus in order to improve the quality of the raw text data (i.e., unannotated text data) in practical applications in machine learning, language data extraction, text processing, technical terms culling, grammatical annotation, information retrieval, machine translation, etc. The resultant output is a moderately edited text corpus, which reduces the amount of workload required in post-editing of texts (Fiser et al. 2012). This implies that pre-editing is a process of adjusting texts before these texts are further processed for subsequent NLP applications. In other words, pre-editing is one such task in which text samples are

N. S. Dash and L. Ramamoorthy, *Utility and Application of Language Corpora*,
https://doi.org/10.1007/978-981-13-1801-6_3

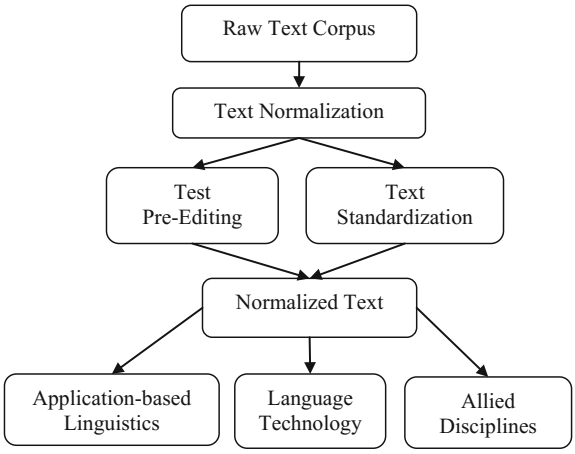

Fig. 3.1 Normalization of text corpus for cross-platform utilization

replotted in certain fixed patterns based on some pre-defined linguistic rules, such as removal of inconsistencies in expression, use of short sentences, avoidance of complex syntactic forms, simplification of ambiguous syntactic structures, redefining patterns of term consistency (Yarowsky 1996). Other pre-editing tasks may involve checking structure of words, marking multiword units, formal consistency verification of larger linguistic units like idioms, phrases, and clauses. All these tasks are required so that further text processing activities (e.g., POS tagging, chunking, parsing, lemmatization) on corpus data become simplified and trouble free (Yarowsky 1994; Xue 2003).

The text normalization process that we describe here includes two major parts: pre-editing and text standardization. Pre-editing, in simple terms, involves several text-based activities like sentence length management, typographic error elimination, punctuation inconsistency removal, header file removal, metadata management, text format simplification, lexical and syntactic ambiguity dissolution, idiomatic expression marking, orthographic style avoidance, non-textual element removal, and domain overlap prohibition (Yeasir et al. 2006). Text standardization process, on the other hand, involves activities like transliteration, grammar checking, tokenization, hyphenation, slash management, period disambiguation, white space management, frozen forms marking, emphatic particle management, indexing, cardinal number management, term usage consistency measurement. Both the parts are so deeply interlinked that occasional overlap of functions across the border is a common phenomenon in text normalization (Sproat et al. 1999, 2001). The theoretical and thematic identity of text normalization can be conceptually perceived from the diagram given below (Fig. 3.1).

3.2 Pre- and Post-editing Trade-off

There is always a trade-off between time and money spent on pre-editing and post-editing for text normalization in document processing. What is important in this enterprise is to keep in mind that if a text document is going to be translated into different target languages, it probably makes sense to spend more time on pre-editing phase than on post-editing. Also, the fact to be remembered is that pre-editing—once it is done in a source text—can solve many issues linked to post-editing (Arens 2004). Post-editing is normally done many times as the number of languages one is processing for translation or POS tagging. Pre-editing, on the other hand, is done normally once in the source text to save strangers lurking at the gate of post-editing. Therefore, the goal of pre-editing is to render the source text in such a way that the quality and output of language processing applications are upgraded. For instance, the outputs of POS tagging and machine translation (MT) will improve in terms of spelling, format of text, grammatical role of lexical items, and overall readability of text if pre-editing is carried out before the text is used as input (Chen and Liu 1992).

In order to do so, it is necessary to distinguish between those rules that improve the quality of the input text and those rules that do not affect the content of the input text. This distinction is necessary to identify the rules that are presented to the user(s) and how these rules are to be actually used for better outputs. This may involve reframing of the whole sentence in the input language at the abstract level in the sense that it should not confuse the language users in understanding the fact that the structure of the revised sentence is a restructured representation of the original input sentence (Chiang et al. 1996).

The basic argument that we advocate in this chapter is that text normalization activities offer many advantages in seamless utilization of a language corpus (Cutting et al. 1992). Therefore, it is essential to render a text corpus in such a manner that the standard of the existing NLP tools that depend on using text corpus as input is improved considerably. In order to do so, it is perhaps necessary to distinguish between the processes that improve the accessibility of texts as well as the processes that keep contents of texts intact. This is required to identify which rules are to be presented to a system and how the results of the rules may be utilized to have better outputs. This may involve reformulation of the structure of the sentence(s) in input texts at the abstract level in a manner that it should not confuse users in utilization of linguistic data and information. The ultimate goal is to create much easier text content within a corpus in respect of its readability of form, accessibility of format, and usability of content.

3.3 Pre-editing and Global Readiness

Global readiness is a process of creating and optimizing the content of a text so that the end users all over the world can grasp its meaning and intention without much

effort (Abel 2011). Based on this proposition, it is fair to visualize global readiness as a multistep process of creating better text by planning, analyzing, and auditing the text for wider application in various domains of NLP. This implies that pre-editing, as a part of global readiness, makes a text ready to play a crucial role in the larger scene of language processing as well as makes a text ready for all sorts of application-based linguistic works.

Translating texts from one language to many languages—either by a man or a machine—takes more time than one language to another. That means, translating texts into multiple languages is a costly proposition on the scale of time, energy, money, effort, and efficiency. It is not that human translators or a translation system is at fault in this enterprise. In reality, both man and machine try to do the best of their ability with the source texts they receive as inputs. The problem lies with the level of complexities involved in the source texts that eventually tells upon the skill of human translators or the robustness of a machine translation system. Keeping this argument in view, in this chapter we argue for creating 'easy text' for both human translators and NLP works including machine translation.

For developing an easy text, it is necessary to focus on the features of a text relating to readability, grammar, format, and reusability. And to achieve these qualities of an input text, one has to adopt several steps of pre-editing. For example, to make corpora of the Indian languages 'global ready' we can adopt the following means as a part of the standardization of the text content:

(a) Management of scientific and technical terms used in texts,
(b) Enforcement of standard grammar rules of the language on texts,
(c) Enforcement of standard stylistic rules on texts for simplicity,
(d) Maintenance of structural consistency of texts,
(e) Elimination of unnecessary words and lexical items from texts,
(f) Shortening of unwisely sentences and segments in texts,
(g) Marking of idiomatic phrases and set expressions in texts.

Since a text often suffers from a variety of naturalization problems, it is rational to streamline and standardize a text through pre-editing to overcome the problems of text processing and computation as well as to make a text global ready for end users. In the following sections, we shall concentrate on various aspects and issues of pre-editing and normalization of texts with reference to examples and instances taken from a Bangla text corpus.

3.4 Pre-editing of Corpus

In general, there are many operations that we can do to make a corpus text global ready. Particularly in the context of a corpus text being used in NLP works, we propose to deploy the following measures on the Indian languages text corpora to make these maximally usable. These pre-editing operations may be practiced in the following areas.

3.4.1 Sentence Management

Keeping sentence length unchanged in the corpus is the optimum priority in pre-editing. There should be no compromise with regard to sentence length. Identification of each sentence as a separate syntactic unit is the most important task. This is necessary to mark and measure the length of a price of text with regard to a number of sentences included in it. Each sentence should be separated from the other if these are combined together. The punctuation marks that are normally used at the end of a sentence should be treated as legitimate sentence terminal markers.

Similarly, it is necessary to identify each segment used in a corpus. Since segments are not sentences, these forms need to be marked separately. The difference between a segment and a sentence should be clearly understood and marked accordingly so that subsequent parsing process that is to be applied to the sentences is not applied over the segments. The structural difference between a segment and a sentence is shown below with examples taken from a Bangla text corpus.

(a) Segment:

(1a) বাংলার লোকসংস্কৃতির সমাজতত্ত্ব।
Bāṅglār lokasaṃskṛtir samājtattva.
"The Sociology of folk culture of Bengal"

(1b) কৃষ্ণনগরের মৃৎশিল্প।
kṛṣṇanagarer mṛt̠silpa.
"Clay art of Krishnanagar"

(b) Sentence:

(2a) বাংলা দেশের আর একটি ঐতিহ্যবাহী প্রাচীন লোক শিল্প হচ্ছে হাতে এবং চাকায় তৈরি মৃৎপাত্র।
Bāṅglā deśer ār ekṭi aitihyabāhī prācīn lok śilpa hacche hāte ebaṃ cākāy tairi mṛtpātra.
"Another old and hereditary folk art of Bengal is hand-made and wheel-made clay plates"

(2b) যন্ত্রানুবাদের মাধ্যমে এক ভাষার লেখ্য তথ্য ও সংবাদ অন্য ভাষায় অনুবাদ করে সারা বিশ্বের সকলের কাছে তাড়াতাড়ি পৌঁছে দেওয়া সম্ভব হবে।
ýantrānubāder mādhyame ek bhāṣār lekhya tathya o saṃbād anya bhāṣāy anubād kare sārā biśver sakaler kāche tāṛātāṛi põuche deoyā sambhab habe.
"Through machine translation, it is possible to reach to everyone in this word quickly by translating written information and data of one language into another language".

It is also necessary to assign unique ID for each sentence (e.g., BNG_FLT_S1990: BNG = Bangla, FLT = Folklore Text, S1990 = Sentence No.: 1990) so that each of the sentences is identified as a separate syntactic unit to be processed independently.

The long sentences which are difficult to read and comprehend should be marked for their unique syntactic structure and properties. In fact, such sentences are mostly ambiguous in nature and are often confusing in meaning. Because of these features, these are quite difficult (if not impossible) to translate into another language.

Finally, verbless sentences may also be marked with the unique flag so that at the time of POS tagging and parsing, special care is taken to address the difficulties involved in such syntactic constructions. Given below are a few Bangla verbless

sentences which require additional care to find their syntactic structure as well as phrases:

(3a) এ সমস্ত তাদের অতুলনীয় কুশলতার প্রমাণ।
e samasta tāder atulanīya kuśalatār pramāṇ.
"These (are) the examples of their unparallel craftsmanship"

(3b) এটি ডোকরা শিল্পের আদি পর্বের প্রথম স্তর।
eṭi ḍokrā śilper ādi parber pratham star.
"This (is) the first phase of the early stage of Dokra Art"

(3c) এই অবস্থায় বর্তমানে ডোকরা শিল্প ও শিল্পী উভয়েই দ্রুত বিলীয়মান।
ei abasthāy bartamāne ḍokrā śilpa o śilpī ubhayei drūta bilīyamān.
"At this present stage, both Dokra Art and artist (are) fast ebbing out"

3.4.2 Typographic Error Elimination

There are several types of typographic error found within words used in a corpus, which is not normalized. With regard to typography, it is possible to classify these errors into five major types as mentioned below with some examples taken from a Bangla text corpus.

(a) **Character Omission**

Here, a particular character from a word is omitted.

হাতা 'hātā'	:	হাত 'hāt'	(ā-allograph is omitted)
ধানী 'dhānī'	:	ধান 'dhān'	(ī-allograph missed)
বাবা 'bābā'	:	ববা 'bbā'	(ā-allograph omitted)
শিশির 'śiśir'	:	শিশর 'śiśr'	(i-allograph omitted)
ইস্কুল 'iskul'	:	ইসুল 'isul'	(consonant k is omitted)

(b) **Character Addition**

In a reverse way, in some cases, a character is added to a word unknowingly.

কমল 'kamal'	:	কমলা 'kamal(ā)'	(ā-allograph is added)
পালকি 'pālki'	:	পালিকি 'pāl(i)ki'	(i-allograph is added)
গামলা 'gāmlā'	:	গামেলা 'gām(e)lā'	(e-allograph is added)
ঘরোয়া 'gharoyā'	:	ঘারোয়া 'gh(ā)royā'	(ā-allograph is added) etc.

(c) **Wrong Character Selection**

In these cases, a wrong character is used within a word in place of a right character.

চাইতে 'cāite'	:	টাইতে 'ṭāite'	('c' and 'ṭ' are closely placed character)
ছাগল 'chāgal'	:	চাগল 'cāgal'	('c' and 'ch' are assigned only one key)
আঙ্গুল 'āṅgul'	:	উঙ্গুল 'uṅgul'	('ā' is changed by 'u')
প্রমাণ 'pramāṇ'	:	প্রমাল 'pramāl'	('ṇ' is changed by 'l')
অথবা 'athabā'	:	অথমা 'athamā'	('b' is changed by 'm')
নিদর্শন 'nidarśan'	:	বিদর্শন 'bidarśan'	('n' is changed by 'b')

(d) **Character Gemination**

In this case, a particular character is doubled due to some technical reasons:

করতে 'karte'	:	কররতে 'karrte'	(r > rr)
বালক 'bālak'	:	বাালক 'bāālak'	(ā > āā)
কলিকাতা 'kalikātā'	:	কললিকাতা 'kallikātā'	(l > ll)
মেয়েলি 'meyeli'	:	মেেয়েলি 'meeyeli'	(e > ee)
লোকটা 'lokṭā'	:	লোককটা 'lokkṭā'	(k > kk)
মহারানি 'mahārāni'	:	মহহারানি 'mahhārāni'	(h > hh)
মাতামহ 'mātāmaha'	:	মাততামহ 'māttāmaha'	(t > tt)

(e) **Character Transposition**

In this case, characters are misplaced in the order of their sequential occurrence in words. The newly formed words are, however, accepted as valid words in the language due. Therefore, this process is known as real word error(RWE) (Chaudhuri et al. 1996).

বালক 'bālak'	:	বাকল 'bākal'	(l...k > k...l)
বদল 'badal'	:	বলদ 'balad'	(d...l > l...d)
কমল 'kamal'	:	কলম 'kalam'	(m...l > l...m)
জমা 'jamā'	:	মজা 'majā'	(j...m > m...j)
কাটা 'kāṭā'	:	টাকা 'ṭākā'	(k...ṭ > ṭ...k)
কপাল 'kapāl'	:	কলাপ 'kalāp'	(p...l > l...p)
মাথা 'māthā'	:	থামা 'thāmā'	(m...th > th...m)
পাশ 'pāś'	:	শাপ 'śāp'	(p...ś > ś...p)

Some examples of non-real word errors, which are also generated through the process of character transposition, are cited below from the Bangla text corpus:

কলকাতা 'kalkātā'	:	কলতাকা 'kaltākā'
সাধারণ 'sādhāraṇ'	:	সাধাণর 'sādhāṇar'
পালিতপুত্র 'pālitaputra'	:	পাতিলপুত্র 'pātilaputra'
হাসপাতাল 'hāspātāl'	:	হাসপালাত 'hāspālāt' , etc.

3.4.3 Punctuation Inconsistency Removal

In principle, the proper use of punctuation marks in the text should be restored. In practicality, it is necessary to have consistent use of punctuation marks in texts. Since some of the symbols work as phrase and sentence boundary markers, their

Table 3.1 Storage of metadata within the header file of a text corpus

Metadata	<Title :: śāmba>, <Language :: Bangla>, <Genre :: Written Text>, <TC :: LIT>, <SC :: Fiction>, <TT :: Imaginative>, <ST :: Book>, <Year :: 1978>, <Edition :: First>, <Volume :: Single>, <Issue :: 0>, <Publisher :: Ananda>, <Place :: Kolkata>, <Author :: kālkuṭ>, <Gender :: Male>, <Age :: 60+>, <Nationality :: Indian>, <Words :: 5120>
Text	মরিতে চাহি না আমি সুন্দর ভুবনে। কথাটা আজ অন্য একটি কথার খেই ধরিয়ে দিল। ধরিয়ে দেওয়া খেই কথাটি অবিশ্যি বিপরীত। না-তে আছে হ্যাঁ। ভ্রমিতে চাহি আমি সুন্দর ভুবনে। <marite cāhi nā āmi sundar bhūbane. kathāṭā āj anya ekṭi kathār khei dhariye dila. dhariye deoyā khei kathāṭi abiśyi biparīt. nā-te āche hyā. bhramite cāhi āmi sundar bhūbane.>

syntactic and functional relevance cannot be ignored. It is, therefore, necessary to check if the requisite use of punctuation mark is present in the text. When two or more sentences are connected without a connector, the proper use of the period (e.g., *full stop*, *question mark*, *exclamation mark*) is absolutely necessary to mark a sentence boundary. Equally important are the proper uses of the comma and other orthographic symbols like '$, &, *' in the text because during POS tagging these are treated as 'residual text elements' and marked accordingly (e.g.,/RD_SYM/,/RD_PUNC/).

3.4.4 Metadata Management

The header file needs to be defined with a text data file in a corpus. The extratextual data and information need to be stored as the metadata in the header file for future reference. The metadata may include name, gender, nationality, and age of author, year of first publication, name of publisher, place of publication, edition used in corpus, type of text. The following table presents a list of items relating to extratextual information of a text and how this kind of information is stored in the metadata of a text file (Table 3.1).

3.4.5 Text Format Simplification

It should be understood that the standard Roman writing conventions like 'italics' and 'underlining' do not work well for many of the Indian languages scripts. In case of the Bangla text, for instance, the process of underlining may not be visually

appealing due to the fact that a number of characters tend to use the space in the lower tier below the baseline. Moreover, the font design of some of the Bangla characters directly affects the usability of this kind of visual effect on texts. Similar argument stands valid in case of use of 'italics' in the Bangla text. Italic writing is not at all appealing for the Bangla font—in both printed and digital formats. Even in case of handwritten texts, the use of italics is the least choice, because this makes a text quite cumbersome in appearance. We can give one or two examples from a Bangla text corpus to show how such uses are very much rare in the language.

Underlined Text

(4a) যোয়ার বপনের পরে ২৫ মিমি বৃষ্টি পেলে যোয়ারের অঙ্কুর খুব ভালো হয়।
(4b) ẏoyār bapaner pare 25 mimi bṛṣṭi pele ẏoyārer aṅkur khub bhālo hay.

(5a) বাঁকুড়ার পোড়া মটির হাতি ও ঘোড়া এখন পৃথিবী বিখ্যাত।
(5b) Bākuṛār poṛā māṭir hāti o ghoṛā ekhan pṛithibī bikhyāta.

Non-underlined Text

(6a) যোয়ার বপনের পরে ২৫ মিমি বৃষ্টি পেলে যোয়ারের অঙ্কুর খুব ভালো হয়।
(6b) ẏoyār bapaner pare 25 mimi bṛṣṭi pele ẏoyārer aṅkur khub bhālo hay.

(7a) বাঁকুড়ার পোড়া মটির হাতি ও ঘোড়া এখন পৃথিবী বিখ্যাত।
(7b) Bākuṛār poṛā māṭir hāti o ghoṛā ekhan pṛithibī bikhyāta.

3.4.6 Ambiguity Dissolution

Ambiguity is a real challenge in text processing. The most sensible suggestion in this case is that it is always better to avoid the use of polysemous words in the text. However, in reality, this is simply impossible as a normal text will invariably have many words which are ambiguous. In this context, the possible suggestion is not to use more than one meaning or grammatical role of a word in the same sentence. But this is also not possible in a natural text since a text user never knows in which sense the word will be accepted by readers. Consider the following examples and ambiguities involved therein:

(8a) ছাত্র হিসেবে আমি আপনাকে বিশ্বাস করি।
(chātra hisebe āmi āpanāke biśvāas kari.)
Reading 1: "As a student I trust you".
Reading 2: "I trust you as a student".

(9a) মুর্খের মত জানতে চেও না।
(murkher mat(å) jānte ceyo nā.)
Reading 1: "Never try to know like a fool".
Reading 2: "Never try to know the opinion of a fool".

(10a) শুনেছি তুমি ভালো কাজ করো।
(śunechi tumi bhālo kāj karo.)
Reading 1: "I have heard that you work well".
Reading 2: "I have heard that you do good works".

Ambiguity is noted not only at the lexical level but at higher level also. In some cases, ambiguity is also noted in a sentence. Structural ambiguity is mostly caused due to the presence of immediately following word (W_2), which, if processed with the preceding word (W_1), may produce a meaning different from their respective independent meaning. That means, an entire sentence can be ambiguous if it is differently interpreted, as some of the examples from a Bangla text corpus show:

(11a) বাচ্চাগুলোকে সাজিয়ে-গুছিয়ে এইমাত্র খেতে বসলাম।
(bāccāguloke sājiye-guchiye eimātra khete baslām.)
1st Reading: "I just sat to eat after dressing the kids"
2nd reading: "I just sat to eat to the dressed kids"

(12a) একশ একটা ফুলের মালা দেবো।
(ekśa ekṭā phuler mālā debo.)
1st reading: "I shall give a garland made of hundred and one flowers"
2nd reading: "I shall give hundred and one flower garlands"

(13a) যেভাবে তুমি ডুবে আছো, সেভাবে আমি ডুবতে পারিনি।
(ýebhābe tumi ḍube ācho, sebhābe āmi ḍubte pārini.)
1st reading: "I cannot plunge as you do"
2nd reading: "I cannot plunge into that emotion where you are"

(14a) পশ্চিমবঙ্গ সরকারের দুগ্ধ বিক্রয় কেন্দ্র।
(paścimbaṅga sarkārer dugdha bikray kendra.)
1st reading: "Milk selling counter of WB Govt."
2nd reading: "Selling counter of WB Govt's milk"

We need to think of some methods through which it is possible to restrict the use of such words or to mark these works with specific notations at the time of pre-editing so that these can solve much confusion about word meanings among the text users in the subsequent use of texts in linguistics and language technology.

3.4.7 Idiomatic Expression Marking

All natural texts are full of set expressions, idiomatic expressions, proverbs, etc. The expressions like *kānā garur bhinna path* (a blind cow has a different path), *jale kumīr ḍāṅgāy bāgh* (crocodile in the water and tiger on the shore), *śāk diye māch ḍhākā* (to hide fish with green vegetables), *marā hātir dām lākh ṭākā* (the price of a dead elephant is one lakh rupees) are quite frequent in use in the texts. People argue that since it is not possible to eliminate idiomatic phrases from a text, it is better to reduce their use as much as it is possible (Raj et al. 2006). In our argument, this is also impossible in natural texts as people are free to use these expressions in texts as they like. It is, therefore, sensible to mark them separately at the time of pre-editing of a text by using chunking method or by some other methods considered suitable for such purposes (Fig. 3.2).

HBT2002	[[prāchīn\JJ samay\N_NN theke\PSP]]_NP [[svāsthya\N_NN]]_NP [[ebang\CC_CCD]]_CCP [[saundarya\N_NN]]_NP [[lābh\N_NN karār\V_VM_VNG janya\PSP]]_NP [[nārkelke\N_NN]]_NP [[bibhinna\JJ bhābe\RB]]_RBP [[byabahār\N_NN karā\V_VM_VNG hayeche\V_VAUX]]_VGF [[।\RD_PUNC]]_BLK

Fig. 3.2 Chunking on a sentence to mark phrase boundary

3.4.8 Orthographic Style Avoidance

It is better to use only Unicode compatible fonts like UTF8 in corpus generation. This solves many problems of text access, management, processing, and utilization. It is always better to use only one font consistently in the corpus, as the use of multiple fonts within a single text may create problems in data processing. The question of capital (upper case) or normal (lower case) fonts is irrelevant in case of texts for the Indian languages scripts, since the Indian language scripts do not follow the system of writing which the Roman script follows.

Similarly, it is better to avoid using word- or character-level styles (e.g., bold, italics, bigger shape of character, striking through word) that may force artificial display of characters in texts, as it is noted in many old and printed texts. Moreover, much care is needed in representation of conjunct characters (e.g., consonant clusters, compound characters), which are made with a combination of several consonant graphemes, vowel allographs, and diacritic symbols (e.g., *ntry, mprs, ṣṭy, pry*). This kind of font combinations may make a text look cumbersome. The task of pre-editing should take care to streamline such orthographic style variations to make a text ready for processing.

Finally, it should be noted that words made with the Roman script should be transliterated into the standard scripts of respective Indian languages, as the following examples show (Table 3. 2).

3.4.9 Non-textual Element Removal

All pictorial or visual elements are to be removed from a text corpus. Similarly, all diagrams, tables, images, graphs, flowcharts, pictures, etc., that are used in printed and digital texts should be removed from a digital text corpus. Mathematical notations, chemical formulae, geometric designs, etc., should also be removed. These elements cannot be translated or tagged in a corpus. It is better not to embed text into images as well as not to embed images into text. A pre-editing process must take care to confirm that pictorial elements can make a digital text corpus 'not-so-user-friendly'

Table 3.2 Original and revised text after orthographic consistency

No	Original text	Revised text
1	এই পার্থক্যের কতকগুলি ঘটেছে জৈবিক (biological) বা বংশগত এবং কতকগুলি সাংস্কৃতিক (cultural) কারণে।	এই পার্থক্যের কতকগুলি ঘটেছে জৈবিক (বায়োলজিক্যাল) বা বংশগত এবং কতকগুলি সাংস্কৃতিক (কালচারাল) কারণে।
	ei pārthakyer katakguli ghaṭeche jaibik (biological) bā baṃśagata ebaṃ katakguli sāṃskṛitik (cultural) kāraṇe.	ei pārthakyer katakguli ghaṭeche jaibik **(bāyolagikyāl)** bā baṃśagata ebaṃ katakguli sāṃskṛitik **(kālcārāl)** kāraṇe.
2	গিরিজনি আলোড়ন ভূ-পৃষ্ঠে অনুভূমিক আকারে (horizontally বা tangential direction) কার্য করিয়া থাকে। ইহাতে ভূ-ত্বকে কোথাও সংনমনের (compression) ফলে সংকোচনের (contraction) অথবা কোথাও টানের (tension) দরুণ প্রসারণের (extension) সৃষ্টি হয়।	গিরিজনি আলোড়ন ভূ-পৃষ্ঠে অনুভূমিক আকারে (হোরাইজন্টালি বা ট্যানজেন্টিয়াল ডিরেকশন) কার্য করিয়া থাকে। ইহাতে ভূ-ত্বকে কোথাও সংনমনের (কমপ্রেশন) ফলে সংকোচনের (কনট্র্যাকশন) অথবা কোথাও টানের (টেনশন) দরুণ প্রসারণের (এক্সটেনশন) সৃষ্টি হয়।
	girijani āloṛan bhū-pṛṣṭhe anubhūmik ākāre (horizontally bā tangential direction) kārya kariyā thāke. ihāte bhū-tvake kothāo saṃnamaner (compression) phale saṃkocaner (contraction) athabā kothāo ṭāner (tension) daruṇ prasāraṇer (extension) sṛṣṭi hay.	girijani āloṛan bhū-pṛṣṭhe anubhūmik ākāre **(horāijanṭāli bā ṭyānjenṭiāl ḍirekśan)** kārya kariyā thāke. ihāte bhū-tvake kothāo saṃnamaner **(kampreśan)** phale saṃkocaner **(kanṭryākśan)** athabā kothāo ṭāner **(ṭenśan)** daruṇ prasāraṇer **(eksṭenśan)** sṛṣṭi hay.

in text processing. Therefore, all such pictorial elements should be removed before a corpus is available for linguistics and language technology works.

3.4.10 Domain Overlap Prohibition

For better access to texts, overlapping in text or subject domains while collecting data for a corpus is not advised (Yarowsky 1994). One has to have precise idea of domains and subdomains during acquisition of language texts. For instance, for many linguistic reasons, poetic texts are removed from a corpus of prose text. Texts collected from one discipline should not be mixed up with texts of other disciplines if not specified and desired beforehand. Similarly, texts obtained from foreign languages including large quotations, statements should be removed from a monolingual corpus. The text corpus, unless otherwise defined and designed, should invariably be monolingual, domain-specific, subject-based, and synchronic (if possible).

3.5 Text Standardization

The most important argument in text standardization is that it should focus on the text to find the issues that may negatively affect the output of a text. It is, therefore, necessary to provide a correction option to improve the quality of a text (Olinsky and

Table 3.3 Words in Roman script and their transliteration in Bangla script

English word	Bangla transliteration	English word	Bangla transliteration
atabrine	ātebrin	cardo	kārḍo
chloroquine	klorokuin	coxa	kaksā
elytra	eliṭrā	femur	phimār
Galia	gyāliyā	lacinia	lyāsiniyā
malaria	myāleriyā	palpifer	pyālpiphār
palp	pyālp	paludrine	pyāluḍrin
plasmochin	plāsmokin	plate	pleṭ
staipes	sṭāipes	tarsus	ṭārsās
tegmina	ṭegminā	tibia	ṭibiyā
trochanter	ṭrokāṇṭār	pneumonia	niumoniyā

Black 2000). It is expected that standardization improves the accessibility of input text to a certain level so that it makes easier to operate processing methods on texts Panchapagesan et al. 2004). It should, however, be kept in mind that depending on language, text standardization rules and strategies may vary.

3.5.1 Transliteration

All technical and scientific terms written in foreign scripts should be transliterated in a text corpus. If possible, the same approach should be adopted for proper names coming from foreign languages, such as English and French names in Bangla text. The process of transliteration should be uniform across all text types in the language so that further discriminations do not arise at the time of named entity recognition or name database generation. For instance, given below is a list of English words and their transliterated forms taken from a Bangla text corpus (Table 3.3).

3.5.2 Grammar Checking

Syntactic errors are commonly found when grammatical concord between subject and predicate is lost within a sentence (Mikheev 2003). The responsibility of a corpus developer is to identify such errors, mark these properly, identify the nature of error, and correct such errors, manually or by rule-based manner. Some examples of grammar correction are presented below, for elucidation.

Wrong form : তিনি সেখানে বসে পড়ল।
tini sekhāne base **paṛla.**

Correct form : তিনি সেখানে বসে পড়লেন।
tini sekhāne base **paṛlen.**
"He (hon.) sat down there"

Wrong form : আমি তখন ক্লাস এইটে পরি।
āmi takhan klās eiṭe **pari.**

Correct form : আমি তখন ক্লাস এইটে পড়ি।
āmi takhan klās eiṭe **paṛi.**
"Then I was reading at class VIII"

Wrong form : তোমার কোনো হেল্প লাগলে আমাকে বলে।
tomār kono help lāgle āmāke **bale.**

Correct form : তোমার কোনো হেল্প লাগলে আমাকে বলো।
tomār kono help lāgle āmāke **balo.**
"Let me know if you need any help"

Wrong form : আমরা আসলে তাকে আমরা অন্তর্দর্শন বলি।
āmrā āsale tāke **āmrā** antardarśan bali.

Correct form : আমরা আসলে তাকে অন্তর্দর্শন বলি।
āmrā āsale tāke (....) antardarśan bali.
Actually, we call it insight

Wrong form : নতুন শাসকবর্গ পুরানো অর্থনৈতিক ভিত্তি মেনে দিয়েছিল
natun śāsakbarga purāno arthanaitik bhittii **mene diyechila**.

Correct form : নতুন শাসকবর্গ পুরানো অর্থনৈতিক ভিত্তি মেনে নিয়েছিল।
natun śāsakbarga purāno arthanaitik bhittii **mene niyechila.**
"The new government accepted the old economic system"

3.5.3 Tokenization

A piece of text, in its raw format, is just a sequence of characters without explicit information about word and sentence boundaries. Before any further processing is done, a text needs to be segmented into words and sentences. This process is known as tokenization. Tokenization divides long character sequences into sentences and sentences into word tokens. Not only words are considered as tokens, but also numbers, punctuation marks, parentheses, and quotation marks are also treated as tokens. Given a sentence, tokenization is the task of chopping it up into small pieces called words (or shorter units, for that matter), with or without inflections. The token is an instance of a sequence of characters in a text that is grouped together as a useful semantic unit (i.e., word) for processing. In alphabetic languages, words are usually

surrounded by white spaces and optionally by punctuation markers or parenthesis or quotes. These elements act as fairly reliable indicators of word or sentence boundaries (Jeffrey et al. 2002).

Tokenization is always language dependent. What is fit for a Bangla text may not be fit for a Hindi text, particularly in case of inflected verb forms. For instance, compare between the Bangla form *ýācchilām* and the Hindi form *ýā rahe the*. Both are single semantic units, but one has a single lexical unit while the other has three separate lexical units grouped together for the same purpose. Therefore, for Bangla *ýācchilām* is a single word with one token, while for Hindi *ýā rahe the* is a single word with three tokens. Given below is a Bangla sentence in its normal and tokenized form.

Normal form: Example Sentence

(8a) যখন মুক চলচ্চিত্রের প্রচলন ছিল, তখন আমরা কোনো ভাষার বন্ধনে আবদ্ধ ছিলাম না।
ýakhan mūk chalacchitrer prachalan chila, takhan āmrā kono bhāṣār bandhane ābaddha chilām nā.
"When the age of silent movie was in vogue, at that time we are not bound with boundary of any language".

Tokenized Form:

যখন (ýakhan)	মুক (mūk)
চলচ্চিত্রের (chalacchitrer)	প্রচলন (prachalan)
ছিল (chila)	তখন (takhan)
আমরা (āmrā)	কোনো (kono)
ভাষার (bhāṣār)	বন্ধনে (bandhane)
আবদ্ধ (ābaddha)	ছিলাম (chilām)
না (nā)	

3.5.4 Hyphenation

Usually, in case of a hyphenated word, hyphen carries the value of a punctuation mark and, therefore, is treated as a separate token. The main purpose of a hyphen is to glue words together. It notifies the reader that two or more elements in a sentence are linked together. Although there are rules and norms governing the use of hyphens, there are situations when we decide whether to add it or not because it can create problems while POS tagging of hyphenated words. For instance, look at the sentence *phaler ojan 70-80 grām paryanta hay* 'The weight of fruit goes up to 70-80 g'. It illustrates the fact that when a hyphen comes in between two sets of words, then a form like '70-80' is to be considered as a single-word unit and it is to be tagged as phaler\N_NN ojan\N_NN 70-80\QT_QTC grām\N_NN parýanta\PSP hay\V_VAUX .\RD_PUNC.

In this case, at least, we miss out an important piece of information of tokenized property or tokens. The string 70-80 is actually two tokens, not one. It separately conveys the idea of having something of the number between 70 and

80. Therefore, instead of keeping them together as one unit, it is better to write them as '70-80'. It constitutes three separate tokens: {70}, {-}, and {80}. Since {-} should be considered as a different token, the sentence should be rewritten as: *phaler ojan 70-80 grām paryánta hay*. There lies a white space between the three tokens mentioned above. In this case, the POS-tagged output will be something like the following: {phaler\N_NN ojan\N_NN 70\QT_QTC -\RD_PUNC 80\QT_QTC grām\N_NN paryánta\PSP hay\V_VAUX .\RD_PUNC}.

The same logic stands valid for the sentence *tāke ei chabir sahakārī-paricālako karā hay* 'He is also made the assistant director of this film.' Keeping in mind the concept of tokenization, the sentence mentioned above should be tagged as the following: {tāke\PR_PRP ei\DM_DMD chabir\N_NN sahakārī\JJ -\RD_PUNC paricālako\N_NN karā\V_VM_VNG hay\V_VAUX .\RD_PUNC}.

There are always some problems that are different from the one mentioned above. For instance, consider the following sentence: *bājir gāner rekarḍiṃer samayi gurudatta ebaṃ gitā rāy eke-aparer kāchākāchi āsen* 'Guru Dutta and Gita Ray got closer to each other during the recording of the songs of Bazi.' Here, the hyphenated word *eke-aparer* should be considered as a single word string made of three tokens including the hyphen in between. But again, *eke-aparer* represents the same meaning as the other form *parasparer* 'to each other' means. Now, the question is how to tag this string in the corpus. In our view, four possible solutions may be considered to overcome this problem:

(a) Break the string *eke-aparer* as three separate tokens as {eke} {–} and {aparer}, and tag them as {eke\PR_PRC} {–\RD_PUND} and {aparer\PR_PRC}.
(b) Since it conveys one meaning, keep it as a single token (*eke-aparer*) and tag it as a reciprocal pronoun {eke-aparer\PR_PRC}.
(c) Remove hyphen and tag the words as two separate reciprocal pronouns, such as {eke\PR_PRC} and {aparer\PR_PRC} as a whole.
(d) Remove the hyphen and tag it as a single reciprocal pronoun, such as {ekeaparer\PR_PRC} as a whole.

The decision has to be taken fast, and it entirely depends on a text annotator. In case of some complex examples such as *do-āś mṛttikā cāṣer janya khub uapayogī* 'Alluvium soil is best suited for farming,' the word with a hyphen mark (i.e., *do-āś*) constitutes a single-word unit. If we break it into two different tokens, the meaning of the word is lost. Hence, it should be tagged as a single common noun {do-āś\N-NN} and not as an adjective and a noun. Similarly, forms like *bren-ṣṭem, tāntrik-tantra, cau-pāyā, spliṭ-brenoyālā, karpās-kalosām* should be tagged either as single-word units or as two-word units as these forms carry a hyphen mark in between the two formative elements.

3.5.5 Slash (/) Problem

In a written text, we sometimes come across forms like '9/10 din' meaning '9/10 days.' This refers to a part of time spanning over 9 or 10 days. It can be tokenized in the following two ways.

(a) If the symbol '/' signifies OR function where the task can be performed in 9 or 10 days, the symbol '/' can be tokenized separately and tagged as a separate punctuation mark: {9\QT_QTC} {/\RD_PUNC} {10\QT_QTC} {din\N_NN}. This is normally done according to a convention adopted in the BIS tagset.
(b) On the other hand, if form *9/10 din* quantifies something as a whole, it should not be tagged as separate units. Rather, it should be tagged as a single composite unit within the category of QT_QTC: {9/10\QT_QTC \dinN_NN}.

Let us look at some other example such as *1/3 aṃśa* '1/3 part.' What should be done with this token? Can we tag it as three separate tokens because collectively they denote a measurement of something and we cannot separate the number 'two-third'? If we do so, it will convey a different concept. We argue that there is no point in tagging as three separate tokens are here. We should take care while we select texts for the corpus, so that it will not create doubts in user's mind if it is an OR separator or it signifies 'a part relationship.' We can create a uniform rule of inference for POS tagging. The POS tagging of the whole text may be as the following:

{gāch\N_NN lāgānor\V_VM_VNG samay\N_NN ālgā\JJ pātāguloke\N_NN bheṅge\V_VM_VNF mūler\N_NN ek tṛtiyāṃśa\QT_QTC keṭe\V_VM_VNF ropaṇ\N_NN karle\V_VM_VNF gāchṭā\N_NN māṭi\N_NN dhare\V_VM_VNF ney\V_VM_VF .\RD_PUNC}

Therefore, during standardization of corpus text, it is always advisable to avoid using '/' whenever it signifies duration. In that case, it will be easy to tag those words.

3.5.6 Period (.) Disambiguation

In English, '.' or period is considered as a punctuation mark that indicates the end of a declarative sentence or statement. In Indian languages, the same function is carried out by *pūrṇacched* 'full stop' ('॥'). The period is also used in Indian language texts, and in most cases, it is not used as a sentence terminal marker, but for some other functions. If a period ('.') appears in an Indian language text, it is mostly used to refer to an abbreviated form of nouns, e.g., *ḍ.*=*ḍāktār* 'doctor,' *sṭ.*=*sṭeśan* 'station,' *gh.*=*ghaṇṭā* 'hour,' *mi.*=*miniṭ* 'minute,' *se.*=*sekeṇḍ* 'second.' In all such cases, a period has one specific function, it is an indicator of the full form of the noun, hence, it should be tagged with the abbreviated form, and both of them should be considered together as a single-word unit. The problem arises when multiple nouns

are abbreviated with recurrent use of period, and all the forms are meant to be put together as a single concept or expression as the following examples show:

Bangla : গতকাল বি.বি.সি থেকে দু জন লোক এসেছিল।
gatakāl bi.bi.si. theke du jan lok esechila.
English: Yesterday two people came from B.B.C.

Bangla: মাটিতে ক্যালসিয়ামের সাত পি.পি.এস. মাত্রা ফলনের জন্য আবশ্যক।
māṭite kyālsiyāmer sāt pi.pi.es. mātrā phalaner janya ābaśyak.
English: Seven P.P.S. doses of calcium are required in the soil for the production.

The question is whether we should treat *bi.bi.si.* (B.B.C.) as a unit of single token or three separate tokens. The rule holds in English that if there is a period (.) within a word, it will not be segmented; instead, it will be treated as a single unit. If this is so, then we should tag this abbreviated form as: {bi.bi.si.\N_NN}. On the other hand, the counter-argument is that the form *bi.bi.si*. 'B.B.C.' is actually made with three abbreviated forms each one of which stands for an independent word: bi. = bṛṭiś (British), bi. = braḍkāṣṭiṃ (Broadcasting) , si. = karporeśan (Corporation). It should, therefore, be treated as three separate entities and tagged accordingly: {bi.\N_NN} {bi\N_NN} {si.\N_NN}. Similarly, in a sentence like *ār. si. boṛāler janma 19-e akṭobar 1903-e ek prasiddha saṅgīt gharānār paribāre hayechila* 'R.C. Boral was born on 19th October 1903 in a highly famous family of musical tradition,' the abbreviated forms *ār. si.* should be treated as two separate words {ār.\N_NNP} and {si.\N_NNP} rather than as a single word {'ār.si.\N_NNP}, as they stand for two proper names (named entities).

3.5.7 White Space

It is necessary to remove the unnecessary white space existing between the words or tokens within a piece of text, as the following examples show.

সহজ সাধ্য (sahaj sādhya)	>	সহজসাধ্য (sahajsādhya)	"easy"
বীজ গুলি (bīj guli)	>	বীজগুলি (bījguli)	"seeds"
পেরে ছিল (pere chila)	>	পেরেছিল (perechila)	"had done"
কথা গুলো (kathā gulo)	>	কথাগুলো (kathāgulo)	"the words"
দিয়ে ছিলেন (diye chilen)	>	দিয়েছিলেন (diyechilen)	"had given"
নরেন কে (naren ke)	>	নরেনকে (narenke)	"to Naren"
মেয়ে দের (meye der)	>	মেয়েদের (meyeder)	"to girls",
উত্তর প্রদেশ (uttar pradeś)	>	উত্তরপ্রদেশ (uttarpradeś)	"Uttar Pradesh" etc.

Since these are single-word units, there is no need to give space in between the two formative parts. On the contrary, it is equally necessary to give proper space between the words where it is needed, e.g., {upakārī.er>upakārī . er} or {ṭāṭkā,lobhanīya>ṭāṭkā, lobhanīya}, {bapankarle>bapan karle}. Here, the two words *ṭāṭkā* and *lobhanīya* are clubbed together, but there should be a space between them because they are two separate words with two different meanings. Therefore,

they should be written separately as *ṭāṭkā* 'fresh' and *lobhanīya* 'attractive.' This implies that typing error with regard to white space should be carefully eliminated from a text corpus. After splitting, these words will stand as independent lexical items with different grammatical functions and meanings, and subsequently these will be tagged under different parts-of-speech.

3.5.8 Emphatic Particles

This is another important text standardization process where emphatic particles need to be properly attached to words. In the existing style of writing in Bangla (and in other Indian languages), emphatic particles are the part of the preceding words and, therefore, should never be written separately. If these are written separately, these should be considered as conjuncts and not as emphatic particles. In the following Bangla examples, particles *-o* and *-i* do not work as conjuncts, but rather work as emphatic particles, and therefore, they should be tagged with their immediately previous words, as shown below:

লাগাইয়া ও ইহা টানা যায়	: লাগাইয়াও ইহা টানা যায়
{lāgāiyā} {o} ihā ṭānā ýāy	: {lāgāiyāo} ihā ṭānā ýāy
যন্ত্রের সাহায্যে ও জমিতে	: যন্ত্রের সাহায্যেও জমিতে
ýantrer {sāhāýye} {o} jamite	: ýantrer {sāhāýyeo} jamite
এই পদ্ধতিতে ই হস্তচালিত	: এই পদ্ধতিতেই হস্তচালিত
ei {paddhatite} {i} hastacālita	: ei {paddhatitei} hastacālita

In the reverse process, it is noted that the conjunct 'o' is sometimes tagged with the previous word as an emphatic particle. This is also a wrong representation of words. In these cases, the conjunct should be detached from the previous word and should be used as a separate lexical item in the text, as the following examples show:

ব্যবধানে ওগভীরতায়	:	ব্যবধানে ও গভীরতায়
byabadhāne {ogabhīratāy}	:	byabadhāne {o} {gabhīratāy}
রামও সীতা	:	রাম ও সীতা
{Rāmo} Sītā	:	{Rām} {o} Sītā
অঙ্গ, বঙ্গও কলিঙ্গ	:	অঙ্গ, বঙ্গ ও কলিঙ্গ
aṅga, {baṅgao} kaliṅga	:	aṅga, {baṅga} {o} kaliṅga

In the above examples, the character 'o' acts as a conjunct. So, it should be separated from its preceding words as well as succeeding words. Since it is a conjunct, it has its own syntactic-cum-semantic function, and thus it should be treated accordingly in the text.

3.5.9 Frozen Terms

In the present Bangla text corpus, there are some forms like HNO_3, H_2SO_3, H_2SO_4, H_3PO_4 which are normally tagged as *Frozen Forms* as these are universally acknowledged as iconic in form and meaning. In the act of text normalization and processing, these forms should remain same for any text of any language. Within this category, we also have mathematical signs (e.g., ×, ÷, +, −, %, /, <, >, =, ∑, Ω), currency symbols (e.g., $, £, ¥, ₹, €), and some specific text symbols (e.g., #, &, @, ©, §, ®, ¢) which should not change in form. They fall into the category of symbol and should be treated in a formal way.

When we come across a character string something like '70%' in Indian language text corpus, we first change the Roman numeral into Indian language numeral and keep the percentage sign (%) separated from the number because this sign carries specific symbolic function and tag. Therefore, the string is taken up as two different tokens and not as a single one.

In text standardization process, we come across another problem relating to symbols such as this: 15:15:15. Since this denotes a relationship of ratio, the symbol ':' carries mathematical information. So, we need to write it in the following format: '15: 15: 15' keeping a space between the digit and the symbol. Similarly, we also come across a string like '6:30:44,' which denotes time indicating hour, minute, and second—all tagged together with the use of the colon (:) between the characters. In this case, also, we have to break the string into three separate units like, '6': '30': '44' and specify that each unit separated by a colon is actually indicating a separate lexical unit with separate meaning and function (though with identical part-of-speech). Alternatively, if the text standardization task is not so rigorous and lexical-bound, one can, for simple comprehension, keep the entire string as an unbroken unit and tag accordingly {\6:30:45\N_NN}.

3.5.10 Indexing

It is noted that most of the Indian languages texts use the Roman numerals in place of standard Indic script numerals. This creates a problem in text processing. To overcome this, we suggest that all the Roman numeral characters should be converted into Indian numeral characters or vice versa. Similarly, all English alphabets should be converted into Indian alphabets at the time of enumeration, for example: (a) = (k), (b) = (kh), (c) = (g), (d) = (gh), (e) = (ṅ).

When we come across digits such as (1) or letters such as (k) within a bracket, they should be treated as single tokens, while the brackets encircling them should be treated as separate symbols. Therefore, it will be better if such strings are marked in the following manners:

(1): {(\RD_PUNC, 1\QT_QTC,)\RD_PUNC}
(k): {(\RD_PUNC, k\QT_QTC,)\RD_PUNC}.

3.6 Conclusion

In this chapter, we suggest that corpus editing and text normalization are necessary for the text corpora of the Indian languages because they offer many advantages in seamless utilization of language texts stored in the corpora (Habert et al. 1998). The goal is to render the source text in such a manner that the existing standard of activities of language technology is considerably improved so that the problems of spelling, format of text, grammatical roles of words, and overall readability of a text, etc., do not create serious hurdles in use of language corpora (Huang et al. 2007).

In order to do so, we need to distinguish between those rules that improve the quality of the input text and those that do not affect the quality of a text in corpus. This distinction should be maintained as it is necessary to identify which rules are presented to the user(s) and how the result of the rules can be used by the text users to have better application outputs. This may even involve reformulation of the whole sentence(s) in the input language at the abstract level in the sense that it should not confuse the text users in the use of language data and texts. The ultimate goal is to create a much easier content within a corpus in respect of its readability of form, accessibility of format, and reusability of content.

References

Abel, S. 2011. Ready for the World: Is Your Content Strategy Global Ready? Blog on 7 April 2011 at: http://thecontentwrangler.com/2011/04/07/ready-for-the-world-is-your-content-strategy-global-ready/.

Arens, R. 2004. A Preliminary Look into the Use of Named Entity Information for Bioscience Text Tokenization. In *Proceedings of the Student Research Workshop (HLT-SRWS'04), HLT-NAACL-2004*, 37–42. PA, USA: Association for Computational Linguistics Stroudsburg.

Chaudhuri, B.B., and U. Pal. 1996. Non-word error detection and correction of an inflectional Indian language. In *Symposium on Machine Aids for Translation and Communication (SMATAC-96)*, New Delhi, April 11–12, 1996 (Hand out).

Chen, K.J., and S.H. Liu. 1992. Word Identification for Mandarin Chinese Sentences. In *Proceedings of the 14th Conference on Computational Linguistics*, 101–107. France.

Chiang, T.H., J.S. Chang, M.Y. Lin, and K.Y. Su. 1996. Statistical Word Segmentation. *Journal of Chinese Linguistics.* 9: 147–173.

Cutting, D., J. Kupiec, J. Pedersen, and P. Sibun. 1992. A Practical Part-of-Speech Tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing*, 133–140.

Fiser, D., N. Ljubesic, and O. Kubelka. 2012. Addressing Polysemy in Bilingual Lexicon Extraction From Comparable Corpora. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012),* Istanbul, Turkey.

Habert, B., G. Adda, M. Adda-Decker, P. Boula de Mareuil, S. Ferrari, O. Ferret, G. Illouz, and P. Paroubek. 1998. Towards Tokenization Evaluation. In *Proceedings of LREC-98*, 427–431.

Huang, C.R., P. Simon, S.K. Hsieh, and Prevot, L. 2007. Rethinking Chinese Word Segmentation: Tokenization, Character Classification, or Wordbreak Identification. In *Proceedings of the ACL 2007 Demo and Poster Sessions*, 69–72. Prague.

Jeffrey, T.C., H. Scuhtze, and R.B. Altman. 2002. Creating an Online Dictionary of Abbreviations from MEDLINE. *Journal of American Medical Informatics Association* 9 (6): 612–620.

Mikheev, A. 2003. Text Segmentation. In *The Oxford Handbook of Computational Linguistics*, ed. R. Mitkov, 201–218. New York: Oxford University Press, Inc.

Olinsky, C., and A. Black. 2000. Non-Standard Word and Homograph Resolution for Asian Language Text Analysis. In *Proceedings of the ICSLP-2000,* Beijing, China, (available: www.cs.cmu.edu/~awb/papers/ICSLP2000_usi.pdf).

Panchapagesan, K., P.P. Talukdar, N.S. Krishna, K. Bali, A.G. Ramakrishnan. 2004. Hindi Text Normalization. In *Presented at the 5th International Conference on Knowledge Based Computer Systems (KBCS),* Hyderabad, India, 19–22 December 2004. (www.cis.upenn.edu/~partha/papers/KBCS04_HPL-1.pdf).

Raj, A., T. Sarkar, S.C. Pammi, S. Yuvaraj, M. Bansal, K. Prahallad, and A. Black. 2006. Text Processing for Text-to-Speech Systems in Indian Languages. In *Proceedings of the ISCASSW6*, 188–193. Bonn, Germany, (www.cs.cmu.edu/~awb/papers/ssw6/ssw6_188.pdf).

Sproat R., A. Black, S. Chen, S. Kumar, M. Ostendorf, and C. Richards. 1999. Normalization of Non-Standard Words: WS'99 Final Report. In *Proceedings of the CLSP Summer Workshop,* Johns Hopkins University, (Available: www.clsp.jhu.edu/ws99/projects/normal).

Sproat, R., A. Black, S. Chen, S. Kumar, M. Ostendorf, and C. Richards. 2001. Normalization of Non-Standard Words. *Computer Speech and Language* 15 (3): 287–333.

Xue, N. 2003. Chinese Word Segmentation as Character Tagging. *Computational Linguistics and Chinese Language Processing.* 8 (1): 29–48.

Yarowsky, D. 1994. Homograph Disambiguation in Text-to-Speech Synthesis. In *Proceedings of the 2nd ESCA/IEEE Workshop on Speech Synthesis*, 244–247. New Paltz, NY.

Yarowsky, D. 1996. Homograph Disambiguation in Text-to-Speech Synthesis. In *Progress in Speech Synthesis*, ed. J.V. Santen, R. Sproat, J. Olive, and J. Hirschberg, 157–172. New York: Springer.

Yeasir, K.M., A. Majumder, M.Z. Islam, N. UzZaman, and M. Khan. 2006. Analysis of and Observations from a Bangla News Corpus. In *Proceedings of the 9th International Conference on Computer and Information Technology (ICCIT-2006),* Dhaka, Bangladesh.

Web Links

http://www.worldcat.org/title/coling-2002.
https://en.wikipedia.org/wiki/Text_normalization.
https://msdn.microsoft.com/en-us/library/ms717050(v=vs.85).aspx.
http://www.cslu.ogi.edu/~sproatr/Courses/TextNorm/.

Chapter 4
Statistical Studies on Language Corpus

Abstract In this chapter, we make attempt to discuss in brief about various types of statistical approaches that are normally used for processing and analyzing a text corpus as well as for obtaining data which may be considered statistically reliable in making some generalized or specific comments on the patterns of occurrence or manner of distribution of linguistic elements in a corpus. Moreover, in this chapter, we try to show how, based on the nature of text used in a corpus, the patterns of quantitative analysis may vary from that of qualitative analysis, although in the long run both types of analysis may be combined together to get a clear picture of the linguistic phenomenon under scrutiny. We also try to give a short history about the use of statistical methods and techniques in the analysis of corpus before and after the introduction of digital corpus as well as describe how descriptive approaches, inferential approaches, and evaluative approaches can be combined together in the act of corpus analysis, linguistics investigation, and inference deduction.

Keywords Statistics · Nature of corpus study · Quantitative analysis
Qualitative analysis · Statistics in corpus study · Descriptive approach
Inferential approach · Evaluative approach · Chi-square test · T-test · ANOVA text
Pearson correlation · Cluster analysis · Factor analysis · Multidimensional scaling

4.1 Introduction

The very idea of putting a natural language within the terrain of twisted quantification is quite eerie. It does not really match with our traditional concept of language analysis and description, which has been mostly descriptive and inferential for ages. For common people like us, linguistics is a subject, which tries to describe the forms and functions of properties of a language, and in this act, it hardly takes any support from Statistics, Mathematics, or other fields of quantification. We have assimilated this philosophy for ages, and based on this philosophy, we have tried to analyze our languages.

N. S. Dash and L. Ramamoorthy, *Utility and Application of Language Corpora*,
https://doi.org/10.1007/978-981-13-1801-6_4

The introduction of various new and recent areas of linguistics, has, however, tried to define language based on mathematical and statistical perspective. Scholars working in the domains like corpus linguistics, computational linguistics, mathematical linguistics, forensic linguistics, stylometrics are now using quantified results of various kinds obtained from corpora in different applications. Also, they are using such quantified results not only for verifying earlier observation but also for making a new observation, formulating new hypotheses, developing language teaching materials and primers, designing tools of language processing, developing systems for language technology, and providing data in language governance. Generally, results obtained from quantitative analysis of corpus present many new insights about a language and its speakers which are not possible to gather from any other manner. This is why scholars like Yule argue that linguists without adequate knowledge of statistical information about various features of a language will make mistake in handling language data as well as in linguistic observation (Yule 1964: 10).

In Sect. 4.2, we propose a dual focus approach which can be the best strategy for applying statistical methods on corpus; in Sect. 4.3, we discuss in some details the nature of corpus study which looks into the processes of both quantitative and qualitative analysis of corpus; in Sect. 4.4, we present a very brief history of use of statistics in corpus study; in Sect. 4.5, we mark the basic approaches to statistical study on corpus which includes descriptive statistical approaches, inferential statistical approaches, and evaluative statistical approaches; and in Sect. 4.6, we highlight the application of statistical results obtained from corpora in linguistics and allied disciplines.

4.2 The Dual Focus Approach

A language corpus, by virtue of its content and composition, can represent numerous varieties of texts which are able to manifest diverse usage varieties of multiple language properties. Because of this unique property, a corpus is a real storehouse of information that can contribute to a large extent for application of various methods of statistical analysis on it. On the other hand, the results obtained from a corpus after application of statistical analysis techniques on it make us richer with new information and insights about a language. A new set of information about a language means a new range of insights about the language for the formulation of new theories and new perspectives for the language.

Keeping this in view in sight, most of the scholars working on corpus linguistics like to adopt the **Dual Focus Approach** in which they are able to combine both the quantitative and the qualitative methods in the study of a corpus (Biber et al. 1998) as well as a language (Fasold 1989). When a corpus (such as the TDIL corpus of Indian languages) is developed with a limited number of samples taken from different types of written texts, it becomes mandatory for a corpus analyst to apply statistical methods in a sensible manner so that features of any kind can be adequately interpreted.

If a corpus is properly representative of a target language, it is considered suitable for making some general observation about the distribution of properties of a language even when the corpus is not fit for large-scale statistical studies. Such small-sized corpora are usually put within the frame of simple quantitative analyses to find out the patterns of use of different linguistic properties in a language as well as to note how these properties are distributed across the text types. The subsequent qualitative analyses on the quantitative results obtained from corpora are often put to a larger population to look at the language from different perspectives. In most cases, the focus is directed toward the analysis of the behavior of linguistic properties to understand their functional roles in language comprehension and application.

In the early years of digital corpus generation, when large-scale corpora like that of the *Bank of English*, the *British National Corpus*, or the *American National Corpus* were not available, quantitative methods were normally applied on small-sized corpora. In those days, a moderately large corpus was enough for making some simple observation about the properties used in a language (Barnbrook 1996; Oakes 1998). And interestingly, most of the observations were found to be true to the linguistic elements when these were compared with the findings obtained from large-scale language corpora.

In general, a language corpus is put to quantitative analysis for various reasons. The frequency of occurrence of various linguistic properties in the corpus is the main reason. This gives us the insight to develop a proper understanding of the patterns of use of different linguistic properties. Moreover, we gather new statistical results to verify earlier counts, modify the counts of distribution of language elements, develop knowledge texts, design technology tools, and describe a language in a more scientific manner.

The qualitative analysis methods do play an equally important role in the quantitative analysis of language data. The qualitative analysis of results obtained from quantitative study often substantiates new findings in different ways to look at the language in a new light. The focus may also be diverted toward the analysis of the behavior of the properties to understand their functional roles in the act of comprehension of a language as well as in utilization.

4.3 Nature of Corpus Study

A language corpus, either big or small, may be subjected to both quantitative and qualitative analyses with the help of various techniques and methods used in statistics. Although the two types of corpus analysis usually form two different perspectives, they are not necessarily incompatible with corpus data. What is important here is the systematic amalgamation of quantitative results with qualitative interpretation in a cross-dependency interface for a better understanding of the linguistic phenomenon observed in a language.

4.3.1 Quantitative Analysis

Quantitative analysis refers to the quantification of data and analysis of data with quantification. In case of quantitative analysis, we intend to classify different types of linguistic elements of a particular language or a language variety based on certain predefined parameters. For instance, in a corpus of written language text samples, we may like to know the frequency of occurrence of orthographic symbols and other characters with an expectation that information of this kind may give us a better understanding about the features of the language as well as supply us useful information for designing primers. By using certain statistical calculation methods, we can try to account for these properties based on which we can construct more complex statistical models in an attempt to explain what is observed and why these are happening. In the long run, these findings may be generalized to a larger population to deduct general views about the language under consideration.

Sometimes, in more specific manner, direct comparisons can be made between two or more corpora using efficient sampling and significance testing techniques to make cross-comparison and cross-reference. Thus, the quantitative analysis allows us to discover which phenomena are likely to be genuine reflections of a language and which are merely chance occurrences. For instance, a quantitative study is noted, which is carried out regarding the frequency of occurrence of words of different part-of-speech in the TDIL Bangla corpus (1995). The results of this analysis give us some interesting information and insight about the language (Table 4.1).

The study (Table 4.1) shows that words belonging to noun category have a much higher frequency of occurrence in the corpus than the words of other categories. In fact, it is exactly equal in percentage in occurrence with words of other categories put together [noun (50%): others (50%)]. The adjective, the second category of the rank, is slightly more than one-third (18%) of the percentage of occurrence of nouns (50%), while words of other categories are far behind in rank with regard to nouns (finite verbs: 11%, postpositions: 9%, non-finite verbs: 4%, adverbs: 4%, pronouns: 3%, and indeclinables: 1%, etc.). This analysis gives us a new kind of information;

Table 4.1 Frequency of use of words of different parts-of-speech

Part-of-speech	TDIL Bangla corpus (1995) (%)
Nouns	50
Adjectives	18
Finite verbs	11
Postpositions	9
Non-finite verbs	4
Adverbs	4
Pronouns	3
Indeclinables	1
Total	100

that is, in Bangla texts, the use of nouns is much higher than other parts-of-speech. This, however, triggers an important question: Why it is so? Why nouns are more frequent in use in texts than words of other parts-of-speech? We have to find out a suitable answer for this.

The main advantage of quantitative analysis is that by just looking at simple frequency of occurrence or distribution of a single language variety, we can construct a precise picture about the patterns of usage of some particular phenomena as well as their relative normality or abnormality in a language. This gives us a clear clue to know a language better as well as to form appropriate questions for making further exploration. In essence, quantitative analysis is a kind of *idealization* of data, because, for some statistical purposes, classification schemes have to be the hard-and-fast type. For instance, in the statistical analysis, it is generally assumed that a particular linguistic item either belongs to category 'x' or it does not. In this kind of classification scheme, the occurrence of linguistic items belonging to a particular class can speak about many hidden aspects of a language, which is otherwise not possible to extract just by looking at the data.

However, in reality, many linguistic properties do not belong to a single category or text type. In fact, they are more consistent with their 'dynamic identity' due to which they can belong across several categories based on some constraints like form, usage, and function which allow them to exist across different categories. For instance, the English word *round* can belong to different parts-of-speech based on its usage in text.

English: round

Round (NN) The match is over after the third round.
Round (ADJ) He has bought a *round* dining table.
Round (IND) He moved it *round* the corner.
Round (FV) I *round* up my discussion with a smile.
Round (Prep) The earth moves *round* the sun.

Therefore, if we say that the word has a high percentage of occurrences in the language; it does not give us any more information than its simple frequency. We are more interested to know its percentage of use as a member of different parts-of-speech as well as its nature of distribution across different parts-of-speech in the language. Only then, we can have a better picture of the word in the language.

Moreover, quantitative analysis tends to sideline rare occurrences of certain linguistic properties, since they do not occur in large number in a particular corpus. Therefore, to ensure that statistical tests provide reliable results, it is essential that minimum frequency information of each linguistic item should be given due importance in subsequent analysis and interpretation. Otherwise, it may force certain finer features of a particular language type to collapse with other features, which, in return, may result in the loss of richness and variety of the language from which the corpus is developed.

4.3.2 *Qualitative Analysis*

Qualitative analysis, on the other hand, aims at providing a complete and sensible description of the observed phenomena which is elicited from quantitative analysis done on a corpus. The primary goal is to justify or describe the phenomenon noted in the quantitative analysis. While quantitative analysis says what it is, the qualitative analysis says why it is. No attempt is made here to assign frequency tags to the linguistic features identified as unique in a corpus. Rather, all rare phenomena receive an equal amount of attention and treatment as more frequent features do.

Qualitative analysis also allows us to draw finer distinctions among the observed phenomena since it is not necessary to shoehorn total set of phenomena within a finite scheme of classification. For instance, the phenomenon of lexical polysemy, an important feature of words, is best recognized by qualitative analysis, since different explicit senses of words, as well as patterns of their sense variation, are analyzed with due importance on every single example noted in a corpus. Thus, all possible sense variations of words are taken into the analysis to make finer distinctions implied by the words when these polysemous words occur in various contexts.

Qualitative analysis is contributive to quantitative analysis. For instance, if we are to analyze the quantitative findings extracted from a corpus (Table 4.1) and to explain why the picture is like this, we have to find out reasons behind such phenomenon behind the patterns of usages of words in the corpus. There might be several reasons behind this. For nouns, there are many proper-named entities (e.g., place names, person names, item names, object names), which are quite frequently used in the corpus. Such nouns are not usually available in a dictionary or in a general lexicon of a language. Moreover, some words, which are usually marked to other parts-of-speech, are found to be used as nouns in the corpus.

A qualitative analysis can also show that many adjectives are used as nouns, and not the reverse one. Moreover, words belonging to noun may also increase in number in corpus because of the entry of many new nouns in the language, a unique feature which does not usually happen for words of other parts-of-speech. It informs us about a unique feature of a language—if a language borrows, it usually does it with nouns, and not with words of other parts-of-speech. This signifies that most of the new words, which have entered into a language, belong to the category of noun. This observation and its subsequent validation by corpus help us decide which words we should consider for inclusion in a dictionary and which words we should ignore. This is one of the most powerful strategies that may be used for developing a corpus-based dictionary of a language.

The main limitation of a qualitative analysis is that the findings cannot be extended to the wider population with the same degree of certainty since most of the findings are not usually tested whether their occurrence is statistically significant or due to chance. To overcome this shortcoming, we can think of adopting a method of combination of the two approaches so that both the approaches can be combined together to contribute toward understanding the phenomenon. This definitely gives us

some advantages to apply combined linguistic information in mainstream linguistics as well as in language engineering.

4.4 Statistics in a Corpus Study: Brief History

The use of statistical estimation processes in language study was widely acknowledged long before the introduction of language corpus in the electronic version. Before the advent of the electronic corpus, scholars have put language to frequency statistics at different points in time, some of which are quite notable. For instance, Miller (1951) initiated a statistical study on the literary style to conduct an information-theoretical analysis on the English language. Herden (1962) made multiple quantitative investigations to observe the pattern of occurrence of characters (i.e., letters and other orthographic symbols) in some English texts. Edwards and Chambers (1964) came out with some interesting results obtained from the application of various statistical methods on the possibilities of occurrence of language properties in English texts. Williams (1940), Dewey (1950), Good (1957), Miller et al. (1958) and many others also made various quantitative investigations on English literary texts. And all these works were done without the use of a corpus in electronic form.

This kind of corpus-based statistical studies is rejuvenated once the corpora in electronic forms are made available to the investigators. We can refer to a few such studies, which are based on electronic corpora of various types. For instance, in an interesting study, Kenny (1982) has used the *Factor Analysis* method to examine whether there is any significant difference between the numbers and patterns of use of words in the writings of three English poets (i.e., Alexander Pope, Samuel Johnson, and Oliver Goldsmith), which could fit into traditional format of heroic couplet.

From a different perspective altogether, Kilgarriff (1996) has used the *Chi-square Test* to examine the linguistic similarities and differences existing among different text corpora of English produced at different points in time by different agencies. Leech et al. (1994) have used the *Log-linear Analysis* on a large corpus of modern English to perform some empirical analyses to identify the non-discrete categories in semantics and to demonstrate the phenomenon of 'semantic gradiance' in word meaning (or lexical semantics). McEnery and Wilson (1996) have also used various statistical methods to analyze both the *Brown Corpus* and the *LOB Corpus* to come out with many new and interesting results to trace finer shades of distinction between the two types of English, namely the *British English* and the *American English*. Biber et al. (1998) have also used various statistical techniques (e.g., *Chi-square Test*, *Multidimensional Scores*, *Factor Analysis*, *T-test*, *ANOVA test*) to study different types of English text corpora they have used in their research and analysis. We have also come across insightful studies where the application of statistical methods on language corpus is discussed in details (Barnbrook 1996; Oakes 1998).

With regard to Indian languages, it is possible to furnish some amount of information since full information is hardly available for reference. For instance, in case of Bangla, although sporadic attempts have been made at the individual level to

record the patterns and frequency of use of various linguistic elements in Bangla, the language has never been put to any kind of frequency-based quantitative analysis with the support of a large and widely representative text corpus. In the history of the study of the language, we have noted that Suniti Kumar Chatterji (1926/1993) has made an attempt to count the frequency of use of words coming from different sources for a Bangla dictionary. He has also made a similar attempt on some selected texts collected from the Old Bangla literature to see how words of different origin are actually distributed in the texts (Chatterji 1926: 241).

After a gap of nearly four decades, Nikhilesh Bhattacharya has made frequency studies on a collection of texts obtained from the writings of a few literary figures of Bengal. The basic goal of this study was to find out how the Bangla words are actually distributed in some of the literary texts composed by great literary figures like Rabindranath Tagore, Sharatchandra Chattopadhyay, Bankimchandra Chattopadhyay, and others (Bhattacharya 1965). On the other hand, Das et al. (1984) have collected some statistical information from a corpus of selected printed documents in Bangla, Assamese, and Manipuri to calculate the frequency of use of characters in these texts. The last on the line goes to Mallik et al. (1998) who have made some quantitative studies on a small collection of sample Bangla texts obtained from various printed text documents to count the frequency of use of letters and words in the language.

All these studies are based on a small collection of text samples, which cannot be claimed to be a corpus in the true sense because these text databases severely lacked in the features of balance, text representation, and largeness. In this regard, the statistical studies presented by Dash (2005) are far more reliable and authentic, since his studies and findings are based on a large corpus of modern Bangla texts containing nearly five million words collected from various subjects and disciplines of language use published between 1981 and 1995. Similar works might have been carried out in other Indian languages, but not reported here due to non-availability of data and information. The most striking point is that after these works, not much progress is made over the decades. Therefore, if one wants to know the basic statistics about the distribution of various linguistic properties in the Indian languages, we have not much information to showcase.

4.5 Approaches to Statistical Study

In an experimental situation, we are presented with some countable events such as the presence of particular linguistic items in a corpus, which are to be measured with relevant information. There are various statistical approaches to achieve this goal, some of which are given in the following:

(a) **Descriptive Statistical Approach**: It enables us to summarize the most important properties of corpus data.

(b) **Inferential Statistical Approach**: It enables us to answer questions related to observed phenomena and formulate a hypothesis about the newly observed varieties.
(c) **Evaluative Statistical Approach**: It enables us to test whether our hypotheses are supported by empirical evidence obtained from the corpus.

Application of all these approaches helps us explore and verify about how mathematical models and theoretical distributions of data are actually related to the reality manifested in the corpus (Oakes 1998: 1). That means the application of various statistical approaches in corpus analysis implies that these have significant roles to play in general language description, language understanding, and language application. In the subsections below, we try to present a brief sketch of each of the approaches stated above with reference to the TDIL Bangla text corpus.

4.5.1 Descriptive Statistical Approach

There are several statistical methods and processes with the descriptive statistical approach. Among these, the 'frequency count statistics' is probably the most straightforward one with which we can work with any kind of quantitative data. In this approach, we first classify all the linguistic items based on some predefined classification schemes. Next, we take an arithmetical count of each of the items which belong to each class of the classification scheme. For instance, we can set up a classification scheme to look at the frequency of occurrence of words belonging to major parts-of-speech in a language: noun (NN), verb (FV), pronoun (PN), adjective (ADJ), adverb (ADV), postposition (PP), and indeclinable (IND). Since words of these classes constitute the major part of the total occurrence of the tokens (i.e., words) in a corpus, it is necessary to know the frequency of use of these words in the corpus to get an idea how these words occur in a language. By using simple frequency count statistics, we can count the number of times words of each part-of-speech have appeared in a corpus, which may be further accumulated in the final list to record patterns of their occurrence in the language.

Although the method of frequency count is quite useful in corpus analysis, we are not happy with simple counts as it has certain disadvantages, particularly when we want to compare one data set with another data set to get better insights. For instance, if we are interested to compare the frequency of occurrence of words in a literary text corpus with that of a mass media text corpus, we cannot address this question by simple statistical counts. Since this method is mostly one dimensional in nature, there is little scope for us for carrying out comparative frequency studies between two or more corpora. Another notable limitation of this method is that it uses a classification scheme which is developed by text investigator, and therefore, it fails to show how items of similar types can be classed into different groups. For instance, in case of word frequency analysis, all the inflected and affixed forms of a single word have to be lemmatized before these words are put to any kind of frequency count.

Since frequency count gives only the number of occurrence of each type of word, it fails to indicate the prevalence of a type in terms of the proportion of a total number of tokens found in a corpus. This is not a problem when two or more corpora, which are put under comparison, are of equal size in a number of tokens. But when they vary in a number of words, any simple frequency count needs to be made with further caution. Even in those situations, where the disparity in size of corpus is not an important issue, it is better to use proportional statistics to present frequencies, since we find them easier to understand than comparing the fractions of unusual numbers. When results are presented in a fraction, or more commonly, in decimal numbers, common language users are often confused with the results and are often misled by their inferences deducted from broken numbers. Therefore, if any result appears to be a very small number, its ratio may be multiplied by hundred to be presented as a percentage with regard to other numbers in the list (see Kennedy 1998).

4.5.2 Inferential Statistical Approach

The inferential statistical approach is considered important for assessing observed phenomena. It is used to find out if observed patterns of use are at all meaningful within the overall study of the linguistic phenomena of a language. For instance, we can use the method of 'significance testing' to find out whether or not a particular finding is a result of a genuine difference between two (or more) linguistic items, or whether it is just due to chance occurrence. For example, we can examine the number of occurrence of the Bangla word *mānuṣ* 'human' in ten different types of text gathered into the Bangla corpus (Table 4.2).

For simplification of analysis, the counts presented in the table include the occurrence of the word in its inflected, affixed, and non-inflected forms. In all text types

Table 4.2 Frequency of use of *mānuṣ* 'human' in different text types of corpus

No.	Different text types	Occurrence of the word
01	Creative writing	196
02	Fine arts	123
03	Social science	178
04	Natural science	92
05	Medical science	152
06	Technology and engineering	45
07	Mass media	70
08	Commerce and industry	30
09	Legal and administration	45
10	Others	64

of the corpus, the word is more often found to be used in its inflected form tagged with various word-formative elements:

(a) Tagged with an enclitic (e.g., *mānuṣṭi* 'the person'),
(b) Tagged with a plural marker (e.g., *mānuṣguli* 'the people'),
(c) Tagged with a case marker (e.g., *mānuṣke* 'to man'),
(d) Tagged with an enclitic and a case marker (e.g., *mānuṣṭike* 'to the man'),
(e) Tagged with a plural marker and a case marker (e.g., *mānuṣgulir* 'of the people').

A simple frequency count of the word in each text type produces the above result (Table 4.2). From the table, it appears that the word is more often used in the creative text followed by texts of social science, medical science, and fine arts. But this cannot be claimed true since we do not know the total word strength of the text included in each text type. Since there is a possibility of variation of a total number of words in each text type, it may happen that the number of words included in the creative text is ten times more than the number of words included in a social science text. In that case, the actual percentage of occurrence of the word in each text type will vary to a large extent. Therefore, to be accurate that this statistics is not just due to coincidence, we need to perform further calculation—a test of statistical significance—to ensure that the quantitative differences among the text types are taken into proper consideration.

There are different significance testing techniques which are used in corpus data analysis based on the type of the variables that are supposed to be compared within a corpus.

(a) **Chi-square Test**: The Chi-square test is one of the most frequently used statistical methods in linguistics. It is normally used to check whether some patterns of observed distribution of words or similar other linguistic elements actually deviate from a uniform distribution of some properties within a language data set (See Greenwood and Nikulin 1996).
(b) **T-test**: The T-test is an important test for statistical significance which is used with interval and ratio level data. It can be used (a) to test whether there are differences between the two groups of data on the same variable, based on the mean (average) value of that variable for each group; (b) to test whether a group's mean (average) value is greater or less than some standard that is already predefined; and (c) to test whether the same group has different mean (average) scores on different variables (See Rice 2006).
(c) **ANOVA Text**: The Analysis of Variance (ANOVA) is a statistical technique that allows us to assess the potential differences in a scale-level dependent variable by a nominal-level variable having two or more categories. Here the observed variance in a particular variable is partitioned into components attributable to different sources of the variation. It provides us a useful statistical test to understand whether or not the mean values of several groups are equal and therefore generalize the t-test to more than two groups. It is useful for comparing (testing) three or more means (groups or variables) for statistical significance (See Rutherford 2001; Cardinal and Aitken 2006).

(d) **Pearson Correlation:** Correlation between sets of data is a measure to know how well these are actually related. The most common measure of correlation is the Pearson Correlation, which is used to measure how strong a relationship exists between the two sets of variables. Pearson correlation is a correlation coefficient which is commonly used in linear regression as it shows the linear relationship between two sets of data. In simple terms, it answers the question: Is it possible to draw a line graph to represent the data? Two letters are used to represent Pearson correlation: Greek letter rho (ρ) for a population, and the Roman letter 'r' for a sample (See Huber 2004; Wilcox 2005, and Katz 2006).

These inferential statistical methods are usually used to measure the differences and similarities between the groups as well as to judge the degree of relationships existing among these variables. However, each technique is usually used separately to produce a test of significance after assessing likelihood to verify if the observed differences could be due to the chance occurrence or if there is any significant interrelation among the observed differences (See Oakes 1998).

4.5.3 Evaluative Statistical Approach

Simple frequency tables taken from a corpus often hide some more general patterns of similarity and difference underlying the linguistic features we are interested in. In principle, these tables can provide us records of differences between particular *samples* (i.e., text type) on particular *variables* (i.e., linguistic features). But they fail to reflect on the pictures of complex interrelationships of similarities and differences concealed within a large number of samples and variables presented in the tables.

In such contexts, we have to depend on the evaluative statistical approaches, which can help us to explain in a far better way why some particular varieties occur in a particular text type and why some particular features behave in the peculiar ways they do in texts. To perform such comparisons, there are many multivariate statistical techniques available (e.g., *Cluster Analysis*, *Factor Analysis*, *Multidimensional Scaling*, *Log-linear Models*), which are regularly used in linguistic research to extract the hidden patterns of relational interfaces of the linguistic features and properties from the raw frequency of data obtained from a corpus.

(a) **Cluster Analysis**

Language properties or conceptually meaningful linguistic elements that share common linguistic characteristics can play important role in our task of analyzing the linguistic elements and describing a language. In this task, we can use a statistical method known as Cluster Analysis to divide language data into groups (clusters) that may be considered meaningful, useful, or both. In language data analysis, the Cluster Analysis process can be highly useful in dividing data into groups (clustering) and assign particular features to the members of this group (classification). In the context of understanding the nature of language data, clusters may be considered as potential

classes and clusters analysis may be treated as a workable technique for finding out and marking the specific classes in overall interpretation of language data as a whole (see Everitt 2011; Manning et al. 2009).

(b) **Factor Analysis**

In linguistic analysis, the Factor Analysis method is normally used to reduce a large number of linguistic variables into fewer numbers of observational factors. By using this technique, we can extract maximum common variance from all variables and put them into a common score for generic analysis. In this model, we can assume several assumptions such as there is a linear relationship among the variables, there is no multicollinearity among the variables, and there is a true correlation between the linguistic variables and factors. In essence, this requires many subjective judgments by the user. Although it is a widely used tool, it can be controversial in certain contexts, because the models, methods, and subjectivity that are used in the variable analysis may be flexible. It is noted that a particular variable may, on certain occasions, contribute significantly to more than one of the components.

(c) **Multidimensional Scaling**

Multidimensional Scaling is considered as an alternative to Factor Analysis in the analysis of the distribution of data (Borg and Groenen 2005: 207–212). In general, the goal of this analysis is to detect meaningful underlying dimensions that may allow us to explain the observed similarities or dissimilarities (distances) between the linguistic objects or features we have been investigating. While through Factor Analysis, we can express the similarities between the objects (e.g., variables) in the correlation matrix; through Multidimensional Scaling, we can analyze any kind of similarity or dissimilarity matrix in addition to the correlation matrices. In general, through Multidimensional Scaling, we attempt to arrange the linguistic variables in a space with a particular number of dimensions so as to reproduce the observed distances. Thus, we can explain the distances in terms of underlying linguistic correlates. In essence, Multidimensional Scaling methods are applicable to a wide variety of linguistic research designs because distance measures between the observable features can be obtained in any number of ways.

The basic aim of these statistical techniques is to summarize a large set of linguistic variables in terms of smaller sets on the basis of some statistical similarities between the original variables available for our analysis. In this process, however, we may lose a negligible amount of information about the differences of the linguistic features (see Biber et al. 1998), but that loss does not usually become vital when we deal with the very large amount of language data with multiple textual variations.

All multivariate statistical techniques generally begin with a process of cross-tabulation of linguistic variables and ample text samples. Scholars like Biber (1993) and Kilgarriff (1996) have used *Factor Analysis* technique to identify relationships between the collocations of some homonymous words in English to investigate their sense differences; McEnery and Wilson (1996) have used methods like *Multidimensional Scaling* to explore the relationships underlying different linguistic variables. Oakes (1998) has used *Log-linear Analysis* to take a standard frequency cross-

tabulation as well as to note which variables seem statistically significant and most responsible for a particular effect generated in the text.

4.6 Conclusion

The idea of putting a natural language within the area of quantification does not really match with our traditional concept of linguistics. Linguistics is an independent subject of investigation and analysis and has hardly encroached within the realm of statistics, mathematics, and quantification. However, the introduction of many new areas of linguistics and allied disciplines (e.g., *computational linguistics*, *corpus linguistics*, *mathematical linguistics*, *lexicography*, *language teaching*, *forensic linguistics*, *stylometrics*, *cognitive linguistics*, *neurolinguistics*) asks for various kinds of quantified results obtained from a corpus for looking at the language from different perspectives and for applying the language in real-life situations. And because of such requirements, the application of various statistical methods and techniques of language corpora has become a part of the present-day linguistic studies.

The results obtained from the application of statistical techniques on corpus have many visible functional relevances. These are used for making observation and hypotheses, developing language teaching materials and primers, as well as designing tools and systems for language technology. In general, the results obtained from quantitative analysis of corpus present many new things that have never been observed before.

References

Barnbrook, G. 1996. *Language and Computers: A Practical Introduction to the Computer Analysis of Language*. Edinburgh: Edinburgh University Press.

Bhattacharya, N. 1965. *Some Statistical Studies of the Bangla Language.* Unpublished Doctoral Dissertation. Indian Statistical Institute, Kolkata.

Biber, D. 1993. Representativeness in Corpus Design. *Literary and Linguistic Computing* 8 (4): 243–257.

Biber, D., S. Conrad, and R. Reppen. 1998. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Borg, I., and P. Groenen. 2005. *Modern Multidimensional Scaling: Theory and Applications*, 2nd ed. Springer-Verlag: New York.

Cardinal, R.N., and M.R.F. Aitken. 2006. *ANOVA for the Behavioural Sciences Researcher*. Mahwah, NJ: Lawrence Erlbaum Associates.

Chatterji, S.K. 1926. *The Origin and Development of the Bengali Language.* Kolkata: Calcutta University Press (Reprinted by Rupa, Kolkata in 1993).

Das, G., S. Bhattacharya, and S. Mitra. 1984. Representing Asamia, Bengali and Manipuri text in Line Printer and Daisy-Wheel Printer. *Journal of the Institution of Electronics and Telecommunication Engineers.* 30 (2): 251–256.

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

Dewey, G. 1950. *Relativ Frequency of English Speech Sounds*. Harvard: Harvard University Press.
Edwards, A.W., and R.L. Chambers. 1964. Occurrence of Various Language Properties in English. *Journal of the Association for Computing Machinery* 2: 465–482.
Everitt, B. 2011. *Cluster Analysis*. Chichester, West Sussex, UK: Wiley.
Fasold, R.W. (ed.). 1989. *Language Change and Variation*. London: John Benjamins.
Good, I.J. 1957. Distribution of Word Frequencies. *Nature* 179: 595.
Greenwood, P.E., and M.S. Nikulin. 1996. *A Guide to Chi-squared Testing*. New York: Wiley.
Herden, G. 1962. *Calculus of Linguistic Observation*. Hague: Mouton & Co.
Huber, P.J. 2004. *Robust Statistics*. New York: Wiley.
Katz, M.H. 2006. *Multivariable Analysis: A Practical Guide for Clinicians*, 2nd ed. Cambridge: Cambridge University Press.
Kennedy, G. 1998. *An Introduction to Corpus Linguistics*. New York: Addison-Wesley Longman Inc.
Kenny, A.J.P. 1982. *The Computation of Style*. Oxford: Pergamon Press.
Kilgarriff, A. 1996. Corpus Similarity and Homogeneity via Word Frequency. In *EURALEX Proceedings*. Gothenburg, Sweden.
Leech, G., B. Francis, and X. Xu. 1994. The Use of Computer Corpora in the Textual Demonstrability of Gradience in Linguistic Categories. In *Continuity in Linguistic Semantics*, ed. C. Fuchs and B. Vitorri, 31–47. John Benjamins: Amsterdam and Philadelphia.
Mallik, B.P., N. Bhattacharya, S.C. Kundu, and M. Dawn. 1998. *Phonemic and Morphemic Frequency in the Bengali Language*. Kolkata: The Asiatic Society.
Manning, C.D., P. Raghavan, and H. Schütze. 2009. *Introduction to Information Retrieval*. Cambridge: Cambridge University Press.
McEnery, T., and A. Wilson. 1996. *Corpus Linguistics*. Edinburgh: Edinburgh University Press.
Miller, G.A. 1951. *Language and Communication*. New York: McGraw-Hills.
Miller, G.A., F.B. Newman, and E.A. Friedman. 1958. Length-Frequency Statistics for Written English. *Information and Control* 1: 370–389.
Oakes, M.P. 1998. *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press.
Rice, J.A. 2006. *Mathematical Statistics and Data Analysis*, 3rd ed. Belmont, CA: Duxbury Press.
Rutherford, A. 2001. *Introducing ANOVA and ANCOVA: A GLM approach*. Thousand Oaks, CA: Sage Publications.
Wilcox, R.R. 2005. *Introduction to Robust Estimation and Hypothesis Testing*. London: Academic Press.
Williams, C.B. 1940. A Note on the Statistical Analysis of Sentence Length as a Criterion of Literary Style. *Biometrika* 31: 356–361.
Yule, G.U. 1964. *The Statistical Study of Literary Vocabulary*. Cambridge: Cambridge University Press.

Web Links

http://math.hws.edu/javamath/ryan/ChiSquare.html.
http://www.ling.upenn.edu/~clight/chisquared.htm.
https://explorable.com/anova.
http://www.socialresearchmethods.net/kb/stat_t.php.
http://www-users.cs.umn.edu/~kumar/dmbook/ch8.pdf.
http://www.yorku.ca/ptryfos/f1400.pdf.
http://documents.software.dell.com/Statistics/Textbook/Multidimensional-Scaling.

Chapter 5
Processing Texts in a Corpus

Abstract In this chapter, we shall make attempt to discuss some of the most common techniques that are often used for processing texts stored in a text corpus. From the early stage of corpus generation and processing, most of these techniques have been quite useful in compiling different types of information about a language as well as formulating some new approaches useful for corpus text analysis and investigation. The important thing is that most of the corpus processing techniques have never been used in language analysis before the advent of the corpus in digital form and application of computer in language data collection and analysis. Another important aspect of this new trend that most of the techniques have strong functional relevance in the present context of language analysis since these techniques give us much better ways to look at the properties of language and utilize them in language analysis and application. We shall briefly discuss some of the most useful text processing techniques such as frequency calculation of characters and words, lexical collocation, concordance of words, keyword in context, local word grouping, and lemmatization. Also, we shall try to show how information and data extracted through the text processing techniques are useful in language description, analysis, and application.

Keywords Corpus · Text processing · Frequency · Characters · Words
Collocation · Concordance · Keywords · Local word grouping · Lemmatization

5.1 Introduction

A digital text corpus, after it is designed and developed, is made available for various processing tasks. There are some well-known text processing techniques which have never been used for analyzing a corpus of earlier years. These techniques are mostly the outcomes of computational interface with language data and are generally used to acquire new insights about language use as well as to shed new lights on the existing methods of language analysis. In reality, application of these techniques on corpus texts results in the finding of new kinds of evidence that may be furnished to

N. S. Dash and L. Ramamoorthy, *Utility and Application of Language Corpora*,
https://doi.org/10.1007/978-981-13-1801-6_5

describe a language as well as its properties from different perspectives with a new interpretation.

There are advantages in the utilization of corpus processing techniques in mainstream linguistic studies, application-based linguistic activities, work of language technology. By applying these techniques, we can gather new sets of data or can collect new types of example to furnish explanations that may fit evidence, rather than adjusting evidence to fit into pre-supposed explanations and inferences. Experience has already taught that processing of a language corpus can produce any evidence and insights, which might directly contradict our intuition-based observation about a language and its properties.

The application of language data and information in the development of tools and systems for language technology has motivated many corpus designers and language engineers to develop different types of corpus processing tools and devices. These include techniques like frequency calculation of characters, words, and sentences; distributional frames of words through concordance; lexical association patterns through collocation; grouping words within local contexts, tracing keyword in specific textual frames, and generating lemmas from inflected and affixed words through lemmatization, etc. Most of these techniques are already developed for English and some other advanced languages (Garside et al. 1987; Souter and Atwell 1993; Thomas and Short 1996; Garside et al. 1997; Oakes 1998; Biber et al.; 1998; Tognini-Bonelli 2001), and nothing worth mentioning is developed for the languages used in India and other so-called less-resourced South Asian languages.

Most of the corpus processing techniques run on text corpus available in digital form. By applying these techniques to a text corpus, we can extract relevant linguistic data and information which are required by the linguists to describe a language or by a language technologist for designing particular systems and devices. Thus, these text processing techniques become quite instrumental in opening up new avenues for language investigation and application that were unknown even a few decades ago.

Keeping this information in the background, in this chapter, we try to discuss some of the most common text processing techniques that are run on corpus with reference to English and some Indian language texts. In Sect. 5.2, we describe frequency counting processes on corpus; in Sect. 5.3, we present a short description on concordance of words; in Sect. 5.4, we discuss the about lexical collocation; in Sect. 5.5, we highlight the aspects of key-word-in-context; in Sect. 5.6, we describe the process of local word grouping; and in Sect. 5.7, we discuss the process of lemmatization.

5.2 Frequency Count

The information about the frequency of occurrence of various language properties within a corpus is of great value in language description as well as in other areas of language study. It is an important piece of information in first language learning (FLL) as it is argued that at the early stage of language learning, a learner should learn first the most frequently used words and their common usage patterns to enhance her

linguistic comprehensibility of the language. In the usage-based scheme of FLL, it is noted that information about the frequency of use of various linguistic items has a strong impact on the overall progress of a language learner (Johns 1991). This leads Barlow (1996) to argue in the following manner:

> …while frequency data is presumably of minor importance in a parameter-setting model of language learning in which the data has only the 'triggering' function in grammar formation, the frequency is very important in an alternative conception of grammar formation based on a model of grammar which is 'use-based.' Such models assume that grammar formation is inductive to a large extent, and that frequency of linguistic usage has a direct effect on the form of the grammar. (Barlow 1996: 5)

At the time of corpus text analysis, information accumulated from the frequency list of words may be rendered in two ways: (a) alphabetical order and (b) numerical order. Moreover, these lists, based on the specific requirement of language users, can again be arranged in ascending and descending orders. There have been questions regarding the relevance of frequency information in the study of a language. To address this question, we can present the following arguments:

(a) A frequency list provides necessary cues to us when we study a piece of text to know how words have occurred in the text with regard to their frequency of occurrence in general.
(b) By examining the frequency list, we gather rudimentary ideas about the basic structure of a language based on which we can plan our future investigation.
(c) Frequency information projects on the patterns of distribution of various linguistic items within a piece of text to understand the content and domain of a text.
(d) Frequency information shades light on overall discourse structure of a text focussing on content, target readership, type, function, addresser, addressee, theme, etc.
(e) Frequency information is a gateway in understanding the ordered development of a text from its simple to complex theme and structure.

The importance of frequency information both in FLL and machine learning can be understood clearly when it is explained how a piece of text is structured in an organized manner. For instance, in case of a technical text, there is a limited number of technical terms that are sparsely distributed at the initial stage of a text. After that, there might be a rush of it, which may be an indicator of a high-level structural boundary in a text. This implies that a text is designed in a systematic fashion as a layman's introduction to a technical subject (Sinclair 1991: 30).

It may happen that two words have the same frequency of use, but while the first one occurs in the first part of a text, the second one occurs at the last part of the text. This kind of distribution is an important clue to a researcher for understanding the structure of a text as well as for furnishing suitable interpretation for this. Such observation helps in selection of texts for FLL in a more pragmatic manner than adopting the traditional methods where a selection of texts is most usually done randomly through intuitions. It is also noted that a randomly selected text may appear

Table 5.1 Alphabetically ordered word list from a Bangla corpus

Bangla word	Percentage (%)	Bangla word	Percentage (%)
nā 'no'	1.153	kintu 'but'	0.423
kare 'doing'	0.989	bā 'or'	0.419
ei 'this'	0.939	karā 'to do'	0.408
o 'and'	0.910	ýā 'that/go'	0.404
hay 'is'	0.764	haye 'being'	0.394
ebaṃ 'and'	0.653	saṅge 'with'	0.371
ýe 'that'	0.653	ek 'one'	0.367
theke 'from'	0.549	kon 'which'	0.358
ār 'and'	0.513	janya 'for'	0.322
tār 'his'	0.497	sei 'that'	0.321

to be an introductory text, but analysis of the frequency of use and distribution patterns in subsequent sections turns it into a typical technical text rich with complex ideas and analysis (Biber and Jones 2009).

In general, frequency-based information is of two types: (a) alphabetic information and (b) numerical information. In true sense, however, these are not two different types of information. Rather, these are two different modes of presentation of same data. In the case of alphabetical frequency information, the listed linguistic items (e.g., *characters, letters*, *morphs*, *words*, *idioms*, *phrases, sentences*) are arranged in alphabetical order to show how items of different alphabetical order are used in a piece of text with a different percentage (Table 5.1).

In numerical frequency list, the same items are arranged according to their degree of occurrence (high to low or vice versa) in the text. The list is formed with a goal to find out which linguistic items are most frequent in use in a text or in a language (Table 5.2). Information is sorted in such a way that the list begins with the most frequent items and continues down to the least frequent items. Generally, in a corpus of hundred thousand words, a frequency list is not much interesting as it fails to say anything convincing. But in a corpus billion of words, a frequency list in numerical order is useful as it provides interesting insights into the language because listings of words in particular order become comparable to the large population of samples for statistical measurement and analysis (Heylen 2005).

Normally, the most frequent items tend to keep suitable distance in distribution, and as a consequence, we can note many marked changes in their order of distribution which become quite significant both in linguistic analysis and generalization. For instance, the 'Asymmetrical Frequency Characteristics' (AFC) of words shows that a small number of very common words make up a high percentage of occurrence in all kinds of texts, while a large number of less used low-frequency words make up the rest (Zipf 1949: 173). This signifies that while words with high frequency may be easily found even in a small-sized corpus, words with low frequency will not occur unless a corpus is made with millions of words obtained from various types of text.

Table 5.2 Numerically ordered word list from a Bangla corpus

Bangla word	Percentage (%)	Bangla word	Percentage (%)
ār 'and'	0.513	tār 'his'	0.497
ei 'this'	0.939	theke 'from'	0.549
ek 'one'	0.367	nā 'no'	1.153
ebaṃ 'and'	0.653	bā 'or'	0.419
o 'and'	0.910	ýā 'go/that'	0.404
karā 'to do'	0.408	ýe 'that'	0.653
kare 'doing'	0.989	saṅge 'with'	0.371
kintu 'but'	0.423	sei 'that'	0.321
kon 'which'	0.358	hay 'becomes'	0.764
janya 'for'	0.322	haye 'being'	0.394

There are many similar linguistic rules and grammatical properties in a language that govern usage of words, multiword units, idiomatic expressions, phrases, collocations, etc. Just as usage, individual senses of words have also definable frequency curves, where more common meanings occur more frequently than less common meanings. In a similar fashion, certain idioms and phrases occur in a language more frequently than others while some set expressions may occur in some specific kinds of text. Similar rules may apply in case of sentences and other linguistic properties used in a language. The basic point is that a frequency calculation method, when it is run on a well-designed and adequately well-represented corpus, generates many new patterns of use of various linguistic items, which was not known before or which are contradictory to our assumptions and expectations.

The information about the frequency of use of various properties in a language carries strong significance both in first and second language education. It is argued that it makes sense to observe the frequency of occurrence before selecting examples of linguistic items for reference and use in grammars and course books for learners (Wills 1990: 142). Also, it becomes useful for lexicographers to know the frequency of use of linguistic items before they take a decision about the selection of lexical entries and others for a dictionary. Thus, information about the frequency of use of items registers some advantages over the intuitive methods of language description, interpretation, and teaching.

The purposes of two types of frequency list are also different. In case of alphabetical frequency list, items are displayed in a tabular form for simple general reference in regular linguistic studies. Thus, it plays a secondary role in the analysis of text, since it is used only when there is a need to check the frequency of particular items in a language. However, it becomes very useful in formulating hypotheses to be tested as well as for checking assumptions made beforehand (Kjellmer 1984).

5.3 Concordance of Words

Technically, concordance program is a process by which one can index words used in a corpus in a separate frame for further analysis. In practice, it is a systematic display of a manageable collection of occurrences of words from a corpus, each instance in its own frame of the contextual environment. That means, in concordance, we index each word with reference to the context of its occurrence in a sentence in the corpus. It is indispensable in the analysis of words because it gives us a wider window to access many important patterns of use as well as its functional and semantic variations in texts (Delorge 2009).

In the earlier years of language analysis, although a technique of this kind was considered useful, we had no scope to apply this since we had no device to the level of a present-day computer by which we could arrange words in the desired manner at a linear level for subsequent observation, analysis, and interpretation. However, the introduction of computer and corpus has made concordance possible to compile and arrange words in the desired fashion for analysis and interpretation. Due to the flexibility of the technique, determination of contextual frames of words is now free based on specific research goal. We can redefine the parameters for selecting words for concordance based on various criteria, such as fixed number words on either side of the target word, finding sentence boundaries of the target words, finding immediate neighboring words of the target word, arranging words based on their root, lemma, base, affixed or inflected forms. What it implies is that the use of concordance in the act of finding general and specific contextual environment of words used in different syntactic frames is an important strategy in understanding the various lexico-grammatical identities of words used in a language.

Although the technique of concordance was not available in the earlier centuries in text analysis, some enterprising scholars diligently did the work manually on some small corpora. They came out with many new findings which inspired others to look at the languages from different perspectives. For instance, the Holy Bible was used as a corpus. It was processed to produce lists of concordance, list of words used, and lists of collocations to prove factual consistency within various parts of the text. Alexander Cruden manually produced concordance lists of some words from an authorized version of the Holy Bible in 1769 (Kennedy 1998: 13). The list of concordance gave scholars new insights into the text to understand it from different angles. After the introduction of the digital corpus, concordance has been quite frequently used not only to study the works of great writers like William Shakespeare (Gibson 1962; Elliott and Valenza 1996), John Milton, James Joyce, T.S. Elliot and others but also to study the texts produced in news media and other domains.

The application of a concordance technique on a text corpus, either small or big, allows us to understand the varieties of linguistic features of a word as well as of the text where the word is used. Concordance opens up many new and innovative ways for studying morphological, lexical, semantic, and syntactic patterns of words as well as the genres and types of a text where the word is used (Barlow 1996). By studying the concordance patterns we can say, with some amount of certainty, about the nature

which these types of food should be	**eaten.**	For most of us it means: eating pleasure
and a wide selection of food will be	**eaten.**	Prepared Softbill food is a good stuff
finger foods and any food that can be	**eaten**	seductively are in! Accomplished fl
are an excellent food and should be	**eaten**	in plentiful quantities. Now to mark
e test with an extract of a commonly	**eaten**	food, we are likely to provoke a pos
casing the amount and variety of food	**eaten.**	Problems could include failure to e
nutrition to the amount and type of food	**eaten,**	the frequency of meals may be an imp
with its body size, whereas the food	**eaten**	by land mammals was not rich enough.
following: Reduce the amount of food	**eaten,**	but not by sacrificing nutritious f
kids. No charge except for the food	**eaten.**	Big fuss made of birthday child. Bu
." Another reason why hot food gets	**eaten**	in hot countries is that chillies an
anyone, or almost everyone, would have	**eaten**	contaminated food. At least that's h
n the name of Scottish food - I have	**eaten**	in places where everything was rolle
ably contain any food you might have	**eaten**	as a snack rather that a proper meal
Bused by overeating it. If a food is	**eaten**	in any form once in three days, or m
ion. If every time a certain food is	**eaten**	the rash becomes worse, or there is
e! It was two days since we had last	**eaten,**	although food had been promised. After
see folic acid. food, substances	**eaten**	or drunk or administered parentally
times happens when the mother hasn't	**eaten**	proper food - I've had a survey don

Fig. 5.1 Concordance of *eaten* in British National Corpus

of distribution, nature of meaning change, nature of inflection or conjugation as well as the patterns of lexical collocation of a word under investigation. We have presented below an example of concordance of the English word *eaten* from the *British National Corpus* to show how the inflected word is used in different syntactic frames to denote different senses based on its syntactic environments (Fig. 5.1).

In language education, a concordance list is a useful resource both for teachers and learners. They can examine the linguistic roles of a word with reference to its distribution in the syntagmatic and paradigmatic axis. With a few sorting operations (left and right sorting), they can refer to a concordance list as a separate text to study if a word is polysemous due to the range of senses embedded within multiple syntactic frames of its use although it shows a single surface form. In the case of data-driven learning (DDL), lessons on grammar and vocabulary far more helpful if these lessons are made with concordance-based word lists compiled from the multitext corpus. In support of this, Johns (1991) has argued that in DDL we can make attempt to remove the middleman in the act of teaching as far as possible and give direct access of the data to the learners so that they can build up their own interpretations with regard to word meanings and uses.

> The assumption that underlies this approach is that effective language learning in itself is a form of linguistic research. The concordance printout offers a unique resource for the stimulation of inductive learning strategies— in particular, the strategies of perceived similarities and differences and of hypothesis formation and testing. (Johns 1991: 30)

In essence, the application of concordance technique on a corpus produces varieties of data and information, which are not possible to gather by any other means. The analysis of such data can challenge the validity of many ideas and observation presented

by scholars about the use of words in texts. Due to this excellent functional relevance, concordance technique is most frequently used in dictionary making, language analysis, defining scientific and technical terms, word sense disambiguation, language teaching, lexicological hypothesis formation, and translation. For instance, a dictionary maker can use concordance technique to search out usage patterns, functions, and sense variations of single words, compound words, idioms, multiword units, and larger word strings from a corpus to be used in a dictionary.

Concordance can be complemented by a range of statistical tools that can provide information about the relative frequency of word use in texts, their distributions across text types, and the list of words with which these forms are most likely to occur or collocate. Therefore, with the help of a concordance, it is no more a difficult task for us to examine all types of occurrence of all kinds of linguistic items in a corpus to describe a language with new information and insights.

5.4 Lexical Collocation

Lexical collocation is technically defined as 'the occurrence of two or more words within a short space of each other in a text' (Sinlciar 1991: 170). This, however, does not reflect the actual operation and its relevance in the study of words in a text. Linguistically, the technique is considered an important method for evaluating relevance and value of consecutive occurrence of two words in a piece of text. It projects into the functional nature of lexical items as well as on the nature of 'interlocking patterns of the lexis' used in a text (Williams 1998). In a simple definition, collocation is a frequently used lexical phenomenon in which two words are quite often found to occur together in a fixed order in various contexts in a language and their co-occurrence is often accepted as correct in the regular practice of language used by people who speak the language (Halliday 1966). This kind of co-occurrence of the two words is not considered as an instance of exocentric compounds.

That means habitual juxtaposition of a word pair with a high frequency of use in the text may be treated as an instance of collocation provided it does not trigger surprise among language users and both the words do not lose their original lexical meanings. Given below a set of English examples to show how the two English words of almost similar senses are having different patterns collocations with different words:

Quick:

Action, answer, change, decision, fix, glance, hand, look, mind, movement, profit, reference, reply, response, shot, succession, time, trip, turn, way, work.

Fast:

Chain, day, end, food, foot, friend, growth, hand, leg, line, look, pace, place, population, rate, speed, step, thing, track, way, world.

In collocation analysis, our primary interest normally centers around the extent to which the actual pattern of these occurrences differ from the patterns that would

have been expected (Barnbrook 1998: 87). We also use collocation to evaluate the argument of cognitive linguistics, which claims that mental lexicon is made up not only of single-word units but also of larger phrasal units, both fixed and variable.

The collocation technique, when it runs on a corpus, produces various kinds of information about the nature of collocation of words. It helps us to understand the position and function of words in a combination with other words in a language. Normally, the lists of citation of contextual use of words obtained from concordance often contain preliminary information about the lexical association of words which we can put to the analysis of collocation of words. A list of collocation may also include some amount of information about the frequency of words used in collocation as well as some sophisticated statistical counts that help us to calculate the combinations needed for comparison and authorization of examples in collocation. If we accept the argument that words are linked with many senses due to their variation of contexts of use, we have to realize that the question of sense variation becomes important in understanding lexical ambiguity, a typical feature of every natural language. It is noted that words exhibit multiple senses by semantic extension of their proto-sense due to the contextual pressure they build up with other words combined in collocation (Delorge 2009).

Lexical collocation is a widely acknowledged linguistic phenomenon of all natural languages. It is addressed in full length with evidence carefully selected from a corpus. For instance, in Bangla, the adjective *kãcā* is used in more than thirty different collocations to denote an equal number of sense variation. Without reference to their frequency and patterns of collocations, it is difficult to understand all the finer senses interlinked with the distributions of the words. In the examples given below (Table 5.3), some sense variations of the word are taken into consideration to make distinctions among the senses implied by the word used in different lexical collocation.

The examples presented above (Table 5.3) signify that with reference to the context of use of words, it is not a difficult task to determine empirically which pairs of words maintain a substantial amount of collocational relation between themselves. The most commonly used formula is Mutual Information of Co-occurrence (MIC) that **compares** the probability of two words occurring together as a regular linguistic phenomenon with a probability of their occurrence as a result of chance. For each pair of words, a statistical count is given where the higher is the score the greater is the degree of collocationality. Thus, MIC helps us in the process of evaluation of the patterns of lexical collocation in a language.

Empirical analysis of the types and patterns of lexical collocation in a language has several advantages:

(a) It helps to extract all double-word units from a corpus and analyze their form and function to be considered for dictionary making, technical translation, and language education.
(b) It helps to group similar collocations of words together to identify their range of sense variation to see if a word (W_1) generates a new sense due to collocation with another word (W_2).

Table 5.3 Variation of sense due to lexical collocation

Word : kā̃cā
POS : adjective
Meaning : "raw"
No. of senses : 30+
Examples :

kā̃cā phal	"Unripe fruit"	kā̃cā māch	"Raw fish"
kā̃cā māṃsa	"Raw meat"	kā̃cā iṭ	"Unburnt brick"
kā̃cā rāstā	"Earthen road"	kā̃cā ghar	"Mud house"
kā̃cā kathā	"Initial talk"	kā̃cā bhāṣā	"Obscene word"
kā̃cā khisti	"Slang word"	kā̃cā sabji	"Green vegetable"
kā̃cā māthā	"Young brain"	kā̃cā lok	"Novice fellow"
kā̃cā hāt	"New hand"	kā̃cā rasid	"Primary draft"
kā̃cā kāj	"Useless work"	kā̃cā raṅ	"Washable color"
kā̃cā sonā	"Pure gold"	kā̃cā cul	"Black hair"
kā̃cā kāṭh	"Wet log"	kā̃cā ojan	"Less weight"
kā̃cā paysā	"Easy money"	kā̃cā ghum	"Incomplete sleep"
kā̃cā bayas	"Immature age"	kā̃cā māl	"Raw material"
kā̃cā mukh	"Filthy mouth"	kā̃cā lekhā	"Poor writing"
kā̃cā kalā	"Green banana"	kā̃cā ẏauban	"Early adulthood"
kā̃cā jal	"Unboiled water"	kā̃cā hisāb	"Initial estimate"

(c) It directs us toward the phenomenon of *semantic gradiance* for conceptualizing how words are able to fabricate new sense by new collocation (Leech et al. 1994).

(d) It helps to identify the contextual differences underlying the use of synonyms. For instance, although the English words *strong* and *powerful* are synonyms, their MIC reveals notable differences. It is noted that *strong* collocates with *motherly, showings, believer, currents, supporter*, *odor*, etc., while *powerful* usually collocates with *neighbor, tool, minority, symbol, figure, weapon, post*, etc. (Church et al. 1991).

(e) It helps to investigate the nature and patterns of grammatical association of two synonymous words. For studying the nature of grammatical association of *little* and *small* in the *British National Corpus* and the *Lancaster-Lund Corpus*, it is found that while *little* co-occurs with concrete, animate nouns (e.g., *little thing(s), little boy(s), little girl(s)*), *small* co-occurs with nouns that indicate quantity (e.g., *small quantity, small amount, small number, small proportion*) (Biber et al. 1998: 94).

(f) It helps to reveal how cultural practice of a speech community shapes up a combination of words in regular use by means of the marked usage of synonymous forms. For instance, in Bangla, the words *din* 'day' and *divas* 'day'

are synonymous, but they vary notably in their patterns of collocation due to some cultural issues and semantic connotations. While *din* mostly occurs with colloquial words having a sense of informality (e.g., *janma din* 'birthday,' *kājer din* 'working day,' *chuṭir din* 'holiday,' *barṣār din* 'rainy day'), *divas* occurs with chaste words having a flavor of formality (e.g., *śramik divas* 'labor's day,' *śiśu divas* 'children's day,' *svādhīnatā divas* 'independence day,' *śahīd divas* 'martyr's day,' *prayāṇ divas* 'salvation day', *mātṛ divas* 'mother's day').

Thus, analysis of collocations can show that words can have important differences in grammatical and lexical association patterns, which interact in an important way resulting differences in the patterns of their distribution across text registers (Evert 2009). In essence, any systematic analysis of examples of collocations collected from corpus can show that nearly synonymous words are rarely equivalent in a sense when considered in terms of their contextual distribution and lexical collocation patterns (Fischer 2000).

In fact, information about the delicate differences underlying lexical collocation patterns between two or more synonymous words has been an important input for the learners in their way of learning a language at an advanced stage. It is useful for the learners to know what kinds of lexical collocation are frequent in use in the language they learn. Also, they should learn to apply these collocations in communication to evoke appropriate impact on listeners. Information of collocation is equally useful in dictionary making, language processing, translation, collocation database generation, and language cognition. However, the most challenging task for non-native speakers is that it is not easy for them to determine which collocations are significant ones to be learned and applied in actual communication events.

5.5 Key-Word-In-Context

Identification and analysis of key-word-in-context (KWiC) is the another technique of corpus processing where the primary goal is to identify the keywords in texts with full reference to their contexts as a part of text display format. KWiC is widely used at the time of processing a text in a corpus to understand the nature and theme of a text with reference to the keywords.

From a technical point of view, the KWiC is the another technique of concordance, which saves a researcher from looking up each and every occurrence of particular words (or a combination of words) in a corpus to study how keywords are used in the construction of a text. However, it differs from concordance in the sense that in case of concordance, a word under investigation is the central focus of attention. In case of KWiC, on the other hand, it is the larger environment or the contextual frame that actually arrests our whole attention. In the first case, it is the word, and in the second case, it is the neighboring words or context that are under scrutiny (Greenacre 2007). Due to this difference in approach, it is better to call it as Context of Keywords (CoK) rather than KWiC.

Generally, in KWiC frame, a keyword is placed at the center of each line with an extra space on either side of the keyword where the length of the context is previously specified by an investigator. The interface displays an environment of two, three, or four words on either side of the keyword located at the center. The format of presentation varies based on the need of a research, as a computer system may be asked to provide relevant citations according to the specifications previously determined. In general, the tasks that a KWiC technique performs are the followings:

(1) Find out all the occurrences of keywords from a corpus, and
(2) Present results in an appropriate format defined by corpus users.

The choice of methods to find out the occurrence of keywords in a corpus is generally made on the basis of processing efficiency of a system. The method of presentation, however, forms a standard display option used in most of the concordance packages. The KWiC registers some advantages over concordance as this display format allows us to select which context to be read to detect the changes noted in a word. In essence, the central block of a display occupied by a keyword keeps a reader's eye in the best position for scanning the lines and noting the contextual frames (Barnbrook 1998: 69).

Due to several display advantages, a KWiC technique adopts various display options (e.g., varied length of KWiC format, sentence context, paragraph context, whole text context) into its system of presentation. The option for variable length format allows the adjustment of the size of the search of text, within which a keyword is entered, in proportion to the size and display facilities provided on the monitor. At sentence- and paragraph-level context, it simply places a keyword in a sentence or a paragraph in which it occurs. The facility to browse all the contexts of a whole text allows moving backward and forward from the point of use of a keyword and permits one to access as much data required for checking the details about the usage of the keyword.

Access to information from a corpus through KWiC technique helps in formulating different objectives in the linguistic description as well as in devising procedures for pursuing these objectives. For instance, the execution of a KWiC program on the *Bank of English* reveals that in English, the most frequently used verbs with reflexive forms are the followings all of which involve 'viewing' as a part of representation or proposition (Barlow 1996).

(a) Find: e.g., *I always **find** myself in trouble.*
(b) See: e.g., *Better **see** yourself.*
(c) Show: e.g., ***Show** yourself the path.*
(d) Present: e.g., ***Present** yourself in the meeting.*
(e) Manifest: e.g., *It was **manifested** in itself*, and
(f) Consider: e.g., *I **consider** myself fortunate.*

Information obtained from the *British National Corpus* through KWiC shows that the verb *manifest* is mostly associated with third person neuter reflexives, whereas the verb *enjoy* occurs with reflexive forms except for the neuter gender (Barlow

1996). Using the same method, the distribution of verbs like *amuse, please, lend, remind*, and others has been studied with examples from the *British National Corpus* to mark their contexts of occurrence. It is observed that most of these verbs are not very common in use and they have a special kind of affinity for reflexive forms.

A KWiC technique is quite useful in understanding the importance of context in analysis of a piece of text, the role of associative words in sense variation, actual behavior of words in context-bound situations, actual environment of occurrence of various linguistic items, and the nature of contextual restrictions exercised in the use of various language properties in speech and writing (Sardinha 1996). The KWiC method is also found convenient and useful at the time of analysis of idiomatic expressions, phrases, and clauses which require additional texts and contexts for better understanding.

Because of many advantages, we like to think KWiC as a text in itself for examining the frequency of words occurring within environments of keywords. The striking advantage of this technique is that linguistic knowledge of the patterns of use of rare lexical items is much useful for moving language learners from intermediate to advanced levels of linguistic proficiency (Fischer 2000). It is not that an entire load of information that is extracted from the contexts is needed at every time, but learners can utilize the information as and when required.

5.6 Local Word Grouping

The method of local word grouping (LWG) is another important way of process corpus for analyzing texts. Unlike concordance and KWiC, this aims at throwing lights from different perspectives on the patterns of use of multiword units, idioms, set phrases, proverbial expressions, and similar linguistic forms in a text. It is noted that LWG technique is useful in those situations where word order is an important criterion for determining the semantic load of sentences, and where a semantic load of an individual word is affected due to the presence of other words.

The LWG technique provides valuable information and insight to deal with the functional behavior of constituents at phrase and clause level at the time of parsing a sentence. The following examples (Table 5.4) taken from the *British National Corpus* show that actual sense of the word *time* cannot be understood properly if the entire length of the LWG is not taken into consideration, because each LWG stands as a unique idiomatic or phrasal expression where each word of the LWG makes contribution in generation of the total sense of the string.

A similar analysis on Bangla text corpus shows that while the Bangla non-finite verbs are mostly followed by finite verbs, nouns are usually followed by postpositions. Thus, usage patterns of Bangla *verb groups* and *noun groups* can be studied by applying an LWG technique on a Bangla corpus. Moreover, these local groups can be analyzed by using local information, which in return, supplies valuable contextual clues for understanding their roles as idiomatic expressions, set phrases, and clausal units.

Table 5.4 LWG of *time* and its usage patterns in English

LWG	Examples
A matter of time	It was only *a matter of time* before this happened
At a time	He took the stairs two *at a time*
At this moment in time	*At this moment in time*, it looks like the business will have to fold
All the time	I get the two of them mixed up *all the time*, they are so similar
At the same time	I can't really explain it, but *at the same time*, I am not convinced
By that time	*By that time* we arrived, the other guests were already there
For the time being	The sale has been canceled *for the time being*
From time to time	This restaurant is pretty good. I come here *from time to time*
In no time	The video has sold fifty thousand copies *in no time*
In the meantime	*In the meantime*, the shares will continue to trade on the open market
Time and again	I have noticed him doing it *time and again*
Time after time	The camera produces excellent results *time after time*
Many a time	*Many a time* they had gone to bed hungry
Many times over	He will live to regret it *many times over*
What kind of time	What *kind of time* are you looking at to get this started?

The information extracted from LWG is equally useful for dissolving lexical ambiguities. Ambiguity may arise from associations of various lexical items within a larger local context (Miller and Leacock 2000: 156). That means understanding the finer shades of meaning is mostly related to the association of specific constituents which combine together to form LWG of different frames and types. This also suggests that the finer shades of meaning are usually conveyed by the internal relations underlying between the constituents along with their distributions in different contextual frames. For instance, it is observed that for many compound words, idioms, and phrases, the meanings that are denoted by a particular association of words cannot be found from meanings of individual words put together. Therefore, for understanding as well as for translation of multiword units, meanings of related words should be grouped together. And this can be derived in a more useful manner from the application of LWG technique on a corpus (Greenacre 2007).

5.7 Lemmatization of Words

The term *lemma* refers to the basic form of a word disregarding its grammatical change it undergoes due to tense and plurality (Biber et al. 1998: 29). Thus, *kill* is a lemma for all the forms made (e.g., *kills, killed, killing, killer*) from it through the application of various processes of affixation and inflection. In case of irregular forms, it is an advantage to put all the morphologically irregular forms under single

lemma for further linguistic analysis. For example, forms like *go, went, gone* can be put under *GO*, while forms like *am, is, are, was, were* should be put under *BE*.

The process of lemmatization involves identification of words which are used in suffixed and inflected forms in texts, removal of suffix or inflection parts, and reducing them to their respective *lemmas*, which are also known as *headwords* that we look for when we look up words in a dictionary. For various works of text processing (e.g., *statistical frequency counts*, *numerical sorting*, *concordance*, *lexical collocation*) lemmatization is a useful method by which we group together different inflected forms of a word, so that, all inflected (and variant forms) are collectively displayed under one *lemma* (Barnbrook 1998: 50).

Lemmatisation is one of the highly useful techniques in corpus-based language research and application. In the area of vocabulary study and dictionary making, it gives helps to produce frequency-based lemma lists as well as distribution information of lemmas for inclusion in the database (Sánchez and Cantos 1997). There are corpora of English and other languages where the process of lemmatization has been used quite successfully to retrieve important lexical data and information for various linguistic works. For instance, the *SUSANNE Corpus* includes lemmatized information for all lexical items, and lemmatized forms are displayed parallel to actual words in a vertical format along with part-of-speech and syntactic information (Beale 1987). Similarly, a part of the *Brown Corpus* contains lemmatized forms of English words along with detailed lexical and grammatical information within a single display format. Also, the *CRATER Corpus of English*, *French*, *and Spanish* (McEnery and Wilson 1996: 43), the *Frankenstein Text* (Barnbrook 1998: 51), the *American National Corpus*, the *Spanish Text Corpus* etc., are lemmatized to identify what kind of inflection the words do have when these are used in texts. Till today, there is not a single lemmatizer available for the Indian languages corpora although scholars have made claims for developing this tool for some of the Indian languages (Sarkar and Bandyopadhyay 2012; Chakrabarty and Garain 2016; Chakrabarty et al. 2016).

There is free software available which may be successfully used for lemmatizing words in a text. However, since most of the software are meant for non-Indian languages, their applicability on Indian language texts is not beyond doubt. On the other hand, since there is no working lemmatizer available that can work for Indian languages, we have no idea how inflected words in the Indian languages can be lemmatized and how these words would behave when the process runs on Indian language texts. To explicate the concept, a standard Bangla sentence is cited below (Table 5.5) where all suffixed and inflected words are arranged in lemmatized forms along with information of their parts-of-speech and meaning.

Through lemmatization, it is possible to extract all possible suffixed and inflected forms of a word from a text and put them under a single lemma. For instance, it is possible for a dictionary maker to assemble together all suffixed and inflected forms of a word searching through the whole corpus and classify patterns of suffixation and inflection accordingly for further actions. If it is done manually, it takes decades to complete a handful of words. Moreover, the task becomes complex, tedious, monotonous, time-consuming, and error-prone.

Table 5.5 Example of lemmatization of words from a Bangla text

Bangla: jātīya rājnītir āṅināy ei phārāk kataṭā sarkārer subidhe bā asubidhe karbe tā rājnaitik biśleṣakerāi balte pārben.

English: The political analysts can only say how this difference will create advantage or disadvantage for the government on the arena of national politics.

Lemma	Surface form	Morphology	POS	Gloss
jāt	jātīya	jāt + īya	ADJ	national
rājnīti	rājnītir	rājnīti + -(i)r	NN	politics
āṅinā	āṅināy	āṅinā + -(ā)y	NN	arena
ei	ei	ei	PN	this
phārāk	phārāk	phārāk	NN	difference
kata	kataṭā	kata + -ṭā	PN	How much
sarkār	sarkārer	sarkār + -er	NN	government
subidhe	subidhe	subidhe	NN	advantage
bā	bā	bā	IND	or
asubidhe	asubidhe	asubidhe	NN	disadvantage
kar	karbe	kar + -be	FV	Will do
tā	tā	tā	PN	that
rājnaitik	rājnaitik	rājnaiti + (-i)k	ADJ	political
biśleṣak	biśleṣakerāi	biśleṣak + -erā + -i	NN	analysts
bal	balte	bal + -te	INF	To say
pār	pārben	pār + -b + -en	FV	Will be able

5.8 Conclusion

The majority of corpus processing tools, systems, and techniques that are available in market or Internet are made for advanced languages like English, Portuguese, French, German, Spanish, Swedish, Dutch, Finnish, Chinese, and Japanese (Biber and Jones 2009; Delorge 2009; Evert 2009; Fischer 2000; Greenacre 2007; Gries 2009; Gries and Divjak 2012; Heylen 2005; Hoffmann 2011; Hox 2010; Johnson 2008; Reif et al. 2013; Wulff 2009). Although these tools and techniques are theoretically applicable to any natural language, in reality, these need to be modified to a large extent before they become useful for the Indian language texts. Modifications are required due to differences existing between the Indian languages in one hand and the other languages on the other hand. Some of the tools may be useful for the Indian language texts if necessary changes and modifications are incorporated into them. Even then, the application of these techniques on the Indian language corpora may not yield expected results due to many linguistic and technical reasons.

Therefore, the best solution is to design indigenous corpus processing tools and systems for the Indian language texts separately. These may differ in approach and methodology adopted for the other languages. The advantage of the systems is that

these are to be designed keeping in view the techniques used for other languages and the basic nature of the Indian language texts. This may lead to the development of better systems as the tools will combine the sophistication of existing techniques with peculiarities of the Indian language texts.

References

Barlow, M. 1996. Corpora for Theory and Practice. *International Journal of Corpus Linguistics* 1 (1): 1–38.

Barnbrook, G. 1998. *Language and Computers*. Edinburgh: Edinburgh University Press.

Beale, A. 1987. Towards a Distributional Lexicon. In *The Computational Analysis of English: A Corpus-Based Approach*, ed. R. Garside, G. Leech, and G. Sampson, 149–162. London: Longman.

Biber, D., and J. Jones. 2009. Quantitative Methods in Corpus Linguistics. In *Corpus Linguistics: An International Handbook*, vol. 2, ed. A. Lüdeling and M. Kytö, 1287–1304. Berlin: Mouton de Gruyter.

Biber, D., S. Conrad, and R. Reppen. 1998. *Corpus linguistics—Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Chakrabarty, A., and U. Garain. 2016. BenLem (a Bengali Lemmatizer) and its Role in WSD. *ACM Trans. Asian and Low-Resource Language Information Processing*. 15(3): 1–12.

Chakrabarty, A., Chaturvedi, A., and U. Garain. 2016. A Neural Lemmatizer for Bengali. In *Presented at 10th Language Resources and Evaluation Conference (LREC)*, Portoroz (Slovenia), May 2016.

Church, K., W. Gale, P. Hanks, and D. Hindle. 1991. Using Statistics in Lexical Analysis. In *Lexical Acquisition*, ed. U. Zernik, 115–164. Englewood Cliff, NJ: Erlbaum.

Delorge, M. 2009. A Diachronic Corpus Study of the Constructional Behaviours of Reception Verbs in Dutch. In *Studies in Cognitive Corpus Linguistics*, ed. B.L. Tomaszczyk and K. Dziwirek, 249–272. Frankfurt: Peter Lang.

Elliott, W., and R. Valenza. 1996. And Then There Were None: Winnowing the Shakespeare claimants. *Computers and the Humanities* 30 (3): 1–56.

Evert, S. 2009. Corpora and Collocations. In *Corpus Linguistics: An International Handbook*, ed. A. Lüdeling and M. Kytö, 1212–1249. Berlin: Mouton de Gruyter.

Fischer, K. 2000. *From Cognitive Semantics to Lexical Pragmatics: The Functional Polysemy of Discourse Particles*. Berlin: Mouton de Gruyter.

Garside, R., G. Leech, and G. Sampson (eds.). 1987. *The Computational Analysis of English: A Corpus-Based Approach*. London: Longman.

Garside, R., G. Leech, and T. McEnery (eds.). 1997. *Corpus Annotation: Linguistic Information from Computer Text Corpora*. London: Addison-Wesley Longman.

Gibson, H.N. 1962. *The Shakespeare Claimants: A Critical Survey of the Four Principal Theories Concerning the Authorship of the Shakespearean Play*. London: Methuen and Co.

Greenacre, M. 2007. *Correspondence Analysis in Practice*, 2nd ed. London: Chapman & Hall.

Gries, S.T. 2009. *Quantitative Corpus Linguistics with R: A Practical Introduction*. London: Routledge.

Gries, S.T., and D. Divjak (eds.). 2012. *Frequency Effects in Language Representation*. Berlin: Mouton de Gruyter.

Halliday, M.A.K. 1966. Lexis as a Linguistic Level. *Journal of Linguistics* 2 (1): 57–67.

Heylen, K. 2005. A Quantitative Corpus Study of German Word Order Variation. In *Linguistic Evidence: Empirical, Theoretical and Computational Perspectives*, ed. S. Kepser and M. Reis, 241–264. Berlin: Mouton de Gruyter.

Hoffmann, T. 2011. *Preposition Placement in English: A Usage-Based Approach*. Cambridge: Cambridge University Press.

Hox, J. 2010. *Multilevel Analysis: Techniques and Applications*, 2nd ed. New York: Routledge.
Johns, T. 1991. Should You Be Persuaded: Two Samples of Data-Driven Learning Materials. *English Language Research Journal* 4: 1–16.
Johnson, K. 2008. *Quantitative Methods in Linguistics*. Oxford: Blackwell.
Kennedy, G. 1998. *An Introduction to Corpus Linguistics*. New York: Addison-Wesley Longman Inc.
Kjellmer, G. 1984. Why 'great': 'greatly', but not 'big': 'bigly'? *Studia Linguistica* 38: 1–19.
Leech, G., B. Francis, and X. Xu. 1994. The Use of Computer Corpora in the Textual Demonstrability of Gradience in Linguistic Categories. In *Continuity in Linguistic Semantics*, ed. C. Fuchs and B. Vitorri, 31–47. John Benjamins: Amsterdam and Philadelphia.
McEnery, T., and A. Wilson. 1996. *Corpus Linguistics*. Edinburgh: Edinburgh University Press.
Miller, G.A., and C. Leacock. 2000. Lexical Representations for Sentence Processing. In *Polysemy: Theoretical and Computational Approaches*, ed. Y. Ravin and C. Leacock, 151–160. New York: Oxford University Press Inc.
Oakes, M.P. 1998. *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press.
Reif, M., J.A. Robinson, and M. Pütz (eds.). 2013. *Variation in Language and Language Use: Linguistic, Socio-cultural and Cognitive Perspectives*. Frankfurt: Peter Lang.
Sánchez, A., and P. Cantos. 1997. Predictability of Word Forms (types) and Lemmas in Linguistic Corpora, A Case Study Based Analysis of the CUMBRE Corpus: An 8-million-word Corpus of Contemporary Spanish. *International Journal of Corpus Linguistics* 2 (2): 259–280.
Sardinha, A.P.B. 1996. Applications of Word-Smith Keywords. *Liverpool Working Papers in Applied Linguistics* 2 (1): 81–90.
Sarkar, S., and S. Bandyopadhyay. 2012. Morpheme Extraction Task Using Mulaadhaar—A Rule-Based Stemmer for Bengali. *JU@FIRE MET 2012*. Working Notes for FIRE 2012 Workshop.
Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
Souter, C., and E. Atwell (eds.). 1993. *Corpus-Based Computational Linguistics*. Amsterdam: Rodopi.
Thomas, J., and M. Short (eds.). 1996. *Using Corpora for Language Research: Studies in the Honour of Geoffrey Leech*. London and New York: Addison Wesley Longman.
Tognini-Bonelli, E. 2001. *Corpus Linguistics at Work*. Amsterdam: John Benjamins.
Williams, G.C. 1998. Collocational Networks: Interlocking Patterns of Lexis in a Corpus of Plant Biology Research Articles. *International Journal of Corpus Linguistics* 3 (1): 151–172.
Wills, J.D. 1990. *The Lexical Syllabus*. London: Collins.
Wulff, S. 2009. *Rethinking Idiomaticity: A Usage-Based Approach*. London: Continuum.
Zipf, G.K. 1949. *Human Behaviour and the Principle of Least Effort: An Introduction of Human Ecology*. Cambridge, Mass.: Addison-Wesley.

Web Links

http://www.ruf.rice.edu/~barlow/corpus.html.
http://www.wordfrequency.info/.
http://www.biblestudytools.com/concordances/.
https://en.wikipedia.org/wiki/Collocation.
http://www.cs.cmu.edu/~ModProb/KWIC.html.
http://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html.

Chapter 6
Corpus as a Primary Resource for ELT

Abstract In this chapter, we argue in favor of teaching English as a second language to the non-native learners with direct utilization of English Language Corpus (ELC). Keeping various advantages of ELC in view, we address here some of the issues relating to the application of ELC as a primary resource of language data and information to be used in the English Language Teaching (ELT) courses for the students who are learning English as a second language. We also discuss here how the learners can access and refer to both speech and text data of ELC in a classroom situation or in a language laboratory for their academic activities. The proposed strategy is meant to be assisted by a computer and based on data, information, and examples retrieved from present-day ELC developed with various text samples composed by native English speakers. The method will be beneficial to the learners if it is used with careful manipulation of tools and techniques used in advanced ELT that advocates utilization of empirical linguistic resources to empower learners. Finally, we argue that the utilization of relevant linguistic data, information, and examples from ELC will enhance the linguistic skills and efficiency of the English learners much better ways than our traditional ELT courses do.

Keywords Interactive ELT · Data-Driven ELT · ELT learners as researchers
Error correction in ELT · Learning sense variation of words in ELT
Learning stylistic variations in ELT

6.1 Introduction

In recent years, English Language Corpora (ELC) of British English or American English are directly utilized as a primary resource in English Language Teaching (ELT) in several non-English countries like China, Japan, Brazil, Saudi Arabia, Malaysia, Singapore, Thailand. In modern ELT courses, the ELC is considered indispensable because they are providing reliable, authentic, and verifiable data, examples, and information that are hardly obtainable by any other means. The most important aspect of this scheme is that these corpora are citing excellent examples of present-day

N. S. Dash and L. Ramamoorthy, *Utility and Application of Language Corpora*,
https://doi.org/10.1007/978-981-13-1801-6_6

English occurring in diverse contexts and varied situations, the reference of which makes learners strongly equipped to absorb finer aspects of language use in varied communicative situations. In fact, the very idea of ELT without reference to ELC appears to be unscientific, because ELCs are giving opportunities to the learners to enrich themselves with present-day English in several directions and diversions. In essence, data and information obtained from ELC provide valuable complementary perspectives toward the traditional linguistic principles of ELT (Biber 1996).

Keeping such advantages in mind, in the following sections of this chapter we address the issues relating to the application of ELC as a primary resource of language data and information in ELT courses for the Indian learners. In Sect. 6.2, we try to justify the rationale behind adopting this new method of ELT; in Sect. 6.3, we introduce the method of interactive ELT where the learners have more active role to play in the process of learning English; in Sect. 6.4, we discuss in brief the basic method of data-driven ELT in which the ELC is the most trusted source of data and information; in Sect. 6.5, we discuss the concept of ELT learners acting as language researchers; in Sect. 6.6, we describe the process of error correction in ELT with reference to ELC; in Sect. 6.7, we discuss how teachers and students can learn sense variation of words with direct reference to ELC; and in Sect. 6.8, we propose how teachers can use various types of English texts to teach stylistic variations to the ELT learners.

In this scheme, the Indian learners are allowed to access and refer to ELC (both speech and text data) directly in the classroom situation or in the language laboratory. They are also allowed to utilize relevant linguistic data, information, and examples from ELC to enhance their linguistic skill and efficiency in English, both in speech and writing. For them, the ELC is a large storehouse of data and information of various types of the modern English language that supply wider empirical perspectives to their mission of learning English.

6.2 The Rationale

There are numerous core domains of first language teaching (FLT) and second language teaching (SLT) where the present-day ELC is utilized to address various language teaching requirements (Botley et al. 2000; Granger et al. 2002). In ELT, the ELC is most often used for the following reasons:

(1) The traditional intuition-based ELT text materials are often found to be misleading, wrong, and fabricated. They contain intuitively invented examples, which normally overlook and ignore the important aspects of usage, and foreground less frequent stylistic choices at the expense of more frequent ones.
(2) The ELC-based ELT materials are more reliable and authentic because these are developed from those corpora which contain data, information, and examples of the real-life use of English. The common choices of linguistic usage are

given more attention and preference than those which are less common and rare in usage (Kübler 2002).

(3) The varieties of examples compiled from present-day ELC are directly used to train the learners about the kinds of English language they will encounter when they are freed to interact in real-life situations.

(4) The citations to the actual use of words, phrases, and sentences derived from the ELC are supplied to the learners to train them about the patterns of use of these items in modern English.

(5) The ELC reveal the range of patterns of English use that the learners should assimilate. They are also presented with frequency information about the use of various linguistic items which become important in lexical choice with regard to learning graded vocabulary in English.

(6) Even within traditional ELT model, the ELC can supply valuable information to learners with regard to the usage patterns of lexical collocation in understanding the patterns of word use in modern English texts (Gavioli 2004).

(7) The ELC empowers the learners to assimilate various issues of English grammar such as the principles that control the use of idiomatic and set expressions, the rules that control the distribution of lexical items and their semantic relations, the network of lexis, and the grammar underlying the surface structure of clauses and sentences.

(8) The ELC makes the advanced learners competent to understand the context-based use of words, set phrases, and idioms as well as help them explore variations of usage of English words and terms across different registers and text types.

(9) The ELC supplies the linguistic and extralinguistic information that contributes toward overall growth of linguistic competence of the learners at primary and advanced levels.

(10) The ELC is useful for looking critically at the existing ELT text materials. Studies showed that there is a considerable amount of difference between what traditional ELT text materials teach and how native speakers actually use the language (Ghadessy et al. 2001).

It has been noted that the advanced ELT textbooks which are based on data taken from ELC usually refer to the most common and frequent choices of linguistic constructions over the rare ones (Hunston 2002: 176). Thus, the direct utilization of information from ELC in ELT radically alters the designing of text materials of ELT and perhaps the discipline as well (Barlow 1996). Moreover, in recent years, we have also noted that there has been an increased interest in corpus-based ELT application for large-scale assessment and in-class instruction (Mukherjee 2002). This has been possible due to the following reasons:

(a) A significant increase in availability of computer and ELC in academic sectors starting from the elementary to the university level.

(b) A notable improvement in ELT methodology for incorporating advanced resources and tools for Natural Language Processing to assess, evaluate, and improve the skill of the learners.

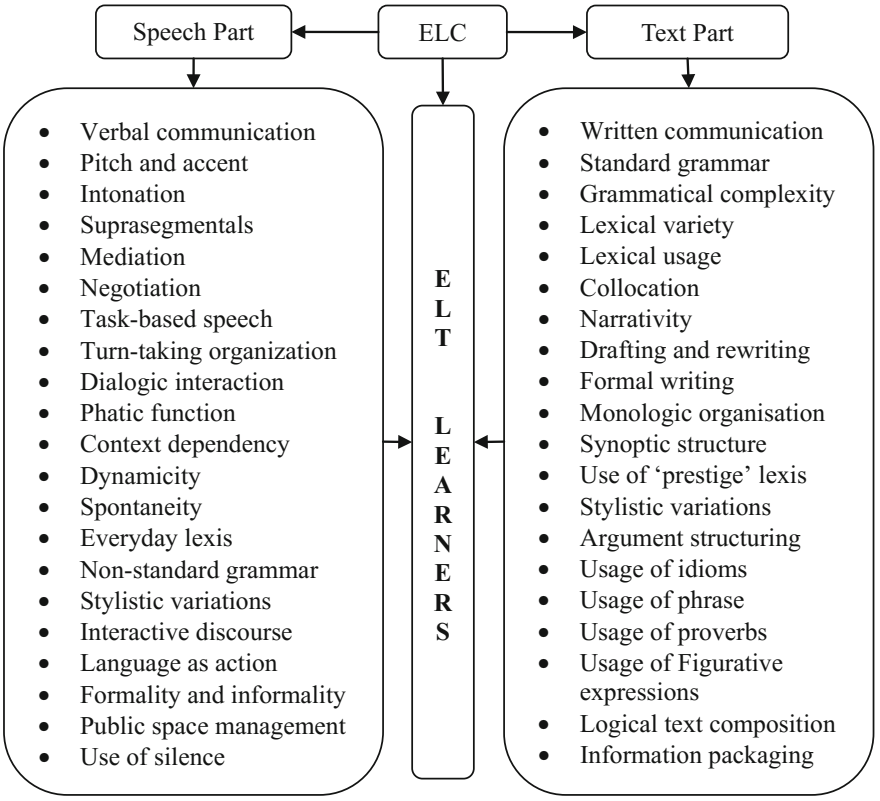

Fig. 6.1 Contribution of ELC for non-native learners in ELT courses

Since these have remained instrumental in helping both the primary and the advanced learners in ELT, the corpus-based approach has shown an inclination to incorporate both English speech and text corpora for input and output work in ELT (Dash 2007) as the following diagram shows (Fig. 6.1).

There are however some limitations with respect to direct utilization of ELC in classroom situations, most of which related to resource manipulation, technical know-how, and trained teacher. Scholars have identified some relevant constraints in direct utilization of ELC in ELT, such as 'the level of experience of learners, paucity of time, curricular requirements, knowledge and skills required of teachers for corpus analysis and pedagogical mediation, and access to resources such as computers and appropriate software tools and corpora, or a combination of these' (McEnery and Xiao 2010). Keeping these limitations in mind, we shall show how ELC may be used as a primary resource in ELT in the following sections.

6.3 Interactive ELT

The ELC directly contributes to computer-assisted interactive ELT, a method that is more effective than our traditional methods of ELT (Kettemann and Marko 2002). In this method, the learners are directly exposed to the speech part of the ELC stored in computer and through an interactive mode they access, utilize, and refer to varieties of speech data to learn the patterns of verbal communication, use of accent in speech, variation of pitch, patterns of intonation, use of various suprasegmental properties like juncture, pause, length, rules of mediation, strategies of negotiation, task-based speech variation, turn-taking organization in speech, dialogic interactions, use of phatic function, use of addressing terms, context-based speech variations, communicative competence, dynamic use of speech, spontaneous production of speech, use of everyday lexis in speech, use of nonstandard grammar, stylistic variations, patterns of interactive discourse, nature of language as action, variation of formal and informal speech, management of public space in speech, and fruitful use of silence in speech. All these aspects are integrated parts of communicative skill of ELT and the best way of learning is to use the speech corpus of the ELC.

On the other hand, the utilization of text data from ELC helps the ELT learners to learn many things of written English, such as method of building up written communication, use of standard grammar in writing, knowledge of grammatical complexity, lexical variety in texts, lexical usage in texts, usage of collocation, narrative variations in text, drafting and rewriting a text, difference between formal and informal writing, monologic organization of texts, synoptic structure of text content, use of 'prestige' lexis, stylistic variations of texts, patterns of argument structuring, usage of idioms, phrases, proverbs and other figurative expressions, process of logical text composition, and the method of information packaging in a text.

Such advantageous usage of ELC makes the learners enthusiastic about English and its properties they are meant to learn, and they probe deep into the texts to explore the intricate patterns of use of various elements in the language. Obviously, their curiosity about the colorful use of language properties is triggered, and they satisfy their curiosity directly from the ELC instead of depending on secondary sources like textbooks, reference books, guidebooks, dictionaries, grammars, and teachers.

Some experiments have been carried in some interactive ELT courses on part-of-speech learning, and these experiments have produced highly encouraging results. The learners are first divided into two main groups: One group is given direct access to ELC while the other group is taught through traditional lecture-based method without reference to ELC. The results obtained from the experiments showed clearly that throughout the course, the learners who are allowed to access ELC by themselves performed far better than the learners taught via traditional lecture-based method (McEnery et al. 1995). In another experiment, it has been observed that the direct exposure to ELC to the advanced learners about the rudiments of the grammar of English yields far better result than those taught through traditional lecture-based method (McEnery and Wilson 1996: 105). The results of such experiments help us argue in favor of using interactive ELT systems for the Indian learners based on ELC.

6.4 Data-Driven ELT

In data-driven ELT, the learners are allowed to act as 'language detectives' under the guidance of instructors (Johns 1997: 101). Theoretically, this approach is based on '3 I's (illustration–interaction–induction) where 'illustration' refers to looking at real-life language data, 'interaction' refers to the act of discussing and sharing opinions and observations, and 'induction' refers to making one's own rule for a particular feature, which 'will be refined and honed as more and more data is encountered' (McEnery and Xiao 2010: 365; Carter and McCarthy 1995: 155). The learners may be allowed to discover themselves the real use of various linguistic items and properties from the ELC while they are learning the language through corpus. Since ELC is a good resource to reveal both the known and unknown linguistic features, the learners can identify many interesting examples and patterns of their own choice, which are not noticed previously, not mentioned in the textbooks or ignored by their instructors. That means the efficiency of the learners will noticeably improve while they search the contexts of linguistic usages through the ELC to deduce meaning and usage patterns of words, idioms, and phrases, etc., used in various English texts.

The strategy may be adopted in this method usually will set up situations in which the learners will be asked to find answers to the questions relating to usage of English words by studying a part of the ELC presented before them in the form of concordance, collocation, or keyword search. Besides, they may be allowed to access the ELC to address regular as well as rare issues of English word usage to develop their efficiency by way of using linguistic information from the ELC. For instance, since most of the learners falter in the correctness of use of *the* in English writing, they may be allowed to explore an ELC (e.g., the *Bank of English*, the *British National Corpus*, the *American National Corpus*, etc.) to find out the most common and the rare usages of *the*. They may also be instructed to isolate sentences where the article is found to be used and classify them according to the functional variation of the article. This will enhance the English skill of the learners in two ways:

(a) They will become inquisitive to explore themselves how the article is used in texts composed by native English speakers and
(b) Errors, if any, they make in their way of interpretation will be verified and corrected with direct reference to the usages in the ELC rather than by their instructors.

Recent research on data-driven ELT has been stressing on encouraging the learners to design their own corpus investigation strategies rather than being monitored by their instructors. If we can apply this strategy then the learners will be allowed to utilize the facility of searching through the ELC by themselves to follow up interesting observations, they may come across in the corpus. This 'discovery learning' (Bernardini 2002) is, however, most suitable for the advanced learners who are trying to fill up the gaps in their knowledge about English rather than laying down the foundation stone in it.

6.5 ELT Learners as Researchers

The training value of ELC is widely recognized in turning the primary 'English language learners' into advanced 'English language researchers.' An example of this method is recently exhibited by Mark Davies at the *Illinois University*, USA. To teach the 'Variation in Spanish Syntax' in class, the advanced students (mostly 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. The learners, as a result of this experiment, came out with many new and unique queries that are never raised or addressed in textbooks or reference materials.

Recent works on ELT have argued that ELC is the best resource to convert primary language learners into advanced language researchers based on the following assumptions (Kirk 2002):

(a) The ELT learners, in some sense, are advanced students, because they have already acquired mastery over their mother tongues or the first languages (L1).
(b) The advanced learners can enhance their linguistic skill in English by themselves by way of direct access to ELC. They hardly need guidance from their instructors.
(c) The advanced learners can carry out their own research on topics of their interest, rather than in-class activities designed by their instructors.
(d) The research works are normally related to advanced issues of syntax and grammar rather than the basic linguistic patterns and usages of English sounds and morphs.
(e) The advanced learners can build up their research works on large ELC rather than depending on small samples of English texts provided by their teachers.

We suggest for following this method for the advanced ELT learners so that they are trained enough in the class with the 'methodologies' for carrying out their own research activities on ELC. The training may involve issues such as learning to carry out a specific linguistic search on ELC, making a hypothesis about the linguistic item(s) in question, testing some hypotheses with examples from ELC, validation or reformulation of earlier hypotheses with reference to the examples taken from ELC. If required, they may be permitted to use several corpora of various length and types along with some search engines to carry out their own search activities on a wide range of topics of their interest on English lexis and grammar. In this case, at least, our main goal is to convert the primary learners into advanced researchers so that they are free to follow the correct methodologies in order to make valid claims about their mastery over the language they are learning.

6.6 Error Correction in ELT

The English texts composed by learners are an important resource for a variety of reasons relating to ELT (Barlow 2000). In fact, systematic analysis of this data can provide reliable evidence to estimate the linguistic skills and efficiencies of the learners as well as to identify the errors and deficiencies they make in the process of their learning. Also, systematic analysis of texts produced by the learners can yield necessary information to redesign the text materials to improve linguistic skills of the default learners as well as to take necessary measures for enhancing their writing and speaking efficiencies. In fact, with this particular goal, the *International Corpus of Learner English (ICLE)* has been generated, which contains extracts of writings produced by the learners of different countries who are learning English as a second language (Granger et al. 2009). The corpus is exhaustively analyzed to know-how the learners acquire efficiency in English or lack of linguistic skills in speaking and writing English. Eventually, their errors are tested and corrected with reference to ELC.

This signifies that within the area of ELT, the corpus that is made with texts produced by the learners can have tremendous potential to be used as useful databases for identifying efficiencies and errors of the learners in learning English. The findings obtained from analysis of such corpora may be used for improving ELT text materials and language teaching techniques as well as for providing necessary remedies to improve English skills of the learners.

6.7 Learning Sense Variation of Words in ELT

Recent corpus-based approaches to ELT have asked for establishing a clear objective criterion toward the semantic study of words. It has been argued that the actual meaning of words used in texts may be derived from the contexts in which they actually occur (Schütze 1997: 142). If this argument is acceptable, then we should make an attempt to find out information regarding the meaning of words from the ELC, which may lead the learners to comprehend meanings of the English words correctly (Mindt 1991).

Traditionally, meanings of English words are taught to the learners based on the information collected from a dictionary or from the intuitive knowledge of English teachers. This strategy has been proved erroneous since the reference to ELC reveals that meaning variations of English words are actually associated with numeral characteristically observable contexts which are not possible to capture unless one refers to a corpus. Not only words and terms, but also compound words, multiword units, idiomatic expressions, and set phrases used in English texts require relevant contextual information for the proper understanding of their meanings.

This signifies that reference to the context of occurrence of English words and other linguistic forms compiled within the ELC may provide the learners with an

An individual is capable of both	great	compassion and great indifference.
Friendship is a	great	the thing in the world.
He lived to a	great	age of his life.
His	great	uncle was a pilot for British Airways.
If you love your life it makes up for a	great	many things you lack.
Love and magic have a	great	deal in common.
Men worry over the	great	a number of diseases.
Rainy season is not a	great	one for travelling.
Sex is a	great	thing provided you know the tricks.
The temptation is a	great	treason: to do right deed for wrong reason.
The universe is a	great	deal bigger than I am.
They do not face a	great	the problem in dealing with them.
This debate made a	great	hole in their friendship.
This has become a	great	survival trick of our species.
To a	great	extent it was the policy we supported.
Our civilization progressed due to the	great	discovery we made.
We had a	great	time at his birthday party.
We must be the	great	arsenal of democracy.
You are a	great	idiot I have ever seen.
You should take	great	care for your infants.

Fig. 6.2 Showing sense variation of 'great' from the BNC through concordance

empirical basis for deciphering finer shades of word meanings. By looking at multi-faceted contexts of use of words in ELC, the learners will be able to understand that a 'fuzzy' concept of word meaning is a better option in sense decipherment because there is no clear-cut boundary between the related meanings of words (Leech et al. 1994). They will also realize that the 'gradiance' of word meanings is actually connected to the frequency of use of particular meaning in a particular context within a piece of text.

It is found that direct reference to ELC can help the learners to understand polysemous nature of many English words that denote multiple senses due to the variation of contexts of their usage. In some recent experiments, it has been shown that the number of sense variations that show up in ELC far exceeds the number of senses provided in English dictionaries Fillmore and Atkins 2000). Also, it has been observed that learners can understand the polysemous nature of English words in a far better way if all the usage variations of words are extracted from ELC and presented before the learners in the form of a concordance list that captures the contexts of their usage (Fig. 6.2). It helps learners to find out multiple sense variations of English words as well as to identify actual senses of words to trace similar words from their mother tongues.

Although the problem of sense discrimination of polysemous English words has been one of the major challenges in ELT for years, the scheme proposed in WordNet gives a good amount of happiness to the learners as it shows how senses are interfaced (Miller et al. 1993). It becomes more useful to them when new figurative senses originate from unknown sources. Since the figurative usage of words is pervasive in a natural language, the core meaning of words becomes almost non-functional

to evoke a new meaning intended by native users (Pustejovsky 1991). In case of corpus-based ELT, this problem is tackled by the following four possibilities:

(a) List up all the sense variations of words noted in corpus in a dictionary of word meaning,
(b) List up frequently used senses of words and employ a generative mechanism to refer to new senses produced by words,
(c) Develop strategy by which learners can identify new meanings and relate these to the original meanings of the words,
(d) Teach learners to analyze words collected in corpora to distinguish literal meanings from the non-literal meanings of words.

Since the figurative use of words is a common feature in English (in other natural languages as well), the strategies proposed above may be useful to a great extent. We, therefore, suggest for identifying figurative senses of words directly from ELC that stores necessary data for the learners to understand the feature. Logically, a direct reference to ELC will illuminate the learners about the basic concepts of literal meaning, metaphor, metonymy, polysemy, and context-sensitive meanings of English words as well as their mutual interfaces that trigger figurative senses. The ELC will provide necessary data and information to the learners to capture the pragmatic factors based on which figurative senses are generated. Moreover, ELC will supply them specific linguistic cues to explore the nature of figurative usage, study their frequency of occurrence in texts, measure the reliability of usage, and evaluate the patterns of semantic generativity of words. Furthermore, ELC will supply the evidence to trace the effect of domains, discourse, and genre in generating figurative senses. Finally, ELC will provide them valuable insight to construct cognitive psychological methods to interpret the figurative senses of words used in English texts.

6.8 Learning Stylistic Variations in ELT

The availability of ELC containing samples of the text of different genres, domains, authors, and media opens up new possibilities for the learners to study stylistic variations in English. In advanced ELT courses, the learners are exposed to individual text types or sample texts written by authors with specific stylistic features. For instance, learners may be trained to find out basic stylistic differences reflected in texts composed by the writers of one generation with that of another generation, while some others may be instructed to find out how the writings of one group of writers stylistically differ from the writings of another group. Such training on comparative stylistic interpretation is possible only when the learners are allowed to access synchronic and diachronic ELC that represents various stylistic features considered relevant in ELT.

Although traditional teaching materials of ELT contain broader issues of stylistics related to genre, a large number of modern ELT materials deal with stylistic features

of specific traits reflected in certain text types. For instance, the learners may be taught to differentiate how English used in newspapers stylistically varies from the English used in scientific texts. Such teachings, however, require relevant corpora from both text types for faithful analysis and observation. The ELC enriched with samples of different texts may be used as a resource for the advanced learners. At the initial stage, the learners are allowed to access the ELC tagged with stylistic features, to begin with, simple general comparisons about the differences in text samples. By way of comparison, they will augment their knowledgebase about stylistic differences manifested in various English texts.

ELC made with different text types are also useful for studying the variation of styles and defining particular styles of writing of a particular author. The learners can use ELC made with the writing samples of a particular author to identify the degree deviation by which the author leans toward various ways of putting linguistic texts (e.g., technical vs. non-technical, choice of vocabulary, long sentences vs. short sentences, the formal vs. informal manner of narration, etc.). This will make them aware not only about the style of writing of a particular author but also understand the styles in which a native English author usually composes the texts.

6.9 Conclusion

In this chapter, we have tried to show how ELC may be used in ELT course for the learners who are learning English as a second language. In a systematic way, we have tried to argue how ELC may be utilized as a resource to be directly accessed by learners in their way of learning English in a classroom situation. The proposed strategy is meant to be assisted by a computer system and is based on data, information, and examples retrieved from the present-day ELC developed with various text samples composed by native English people. The strategy will be beneficial to the learners if used with careful manipulation of tools and techniques used in advanced ELT courses that advocate for the utilization of empirical linguistic resources to empower learners. This corpus-based approach becomes highly suitable in the present situation of India where English is used as a 'lingua franca.'

In recent years, the introduction of ELC in ELT course has made a remarkable breakthrough that leads toward modification of systems and resources traditionally used in ELT. But the most unfortunate thing is that, in comparison with other, Indian learners lag far behind in corpus-based ELT method. We have never come across a situation where Indian learners are taught through corpus-based method. Since there is no scope for speculation about the application relevance of ELC in ELT, the time has come to redirect our attention toward this new approach to rejuvenate the discipline with new lease of life. The learners will not only benefit from this method but also will excel in English when they are put into competition with fellow competitors from other countries. However, before we adopt this method we should make attempt to procure large corpora of British English and American English to be used in ELT courses.

References

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 *Multilingual Corpora in Teaching and Research*, ed. S.P. Botley, A. McEnery, and A. Wilson, 106–115. Amsterdam-Atlanta, GA: Rodopi.

Bernadini, S. 2002. Exploring New Directions for Discovery Learning. In *Teaching and Learning by Doing Corpus Analysis*, ed. C.B. Kettemann and G. Marko, 42–51. Amsterdam-Atlanta, GA: Rodopi.

Biber, D. 1996. Investigating Language use Through Corpus-Based Analyses of Association Patterns. *International Journal of Corpus Linguistics* l (2): 171–198.

Botley, S.P., A. McEnery, and A. Wilson (eds.). 2000. *Multilingual Corpora in Teaching and Research*. Amsterdam-Atlanta, GA: Rodopi.

Carter, R., and M. McCarthy. 1995. Grammar and the Spoken Language. *Applied Linguistics* 16 (2): 141–158.

Dash, N.S. 2007. *Language Corpora and Applied Linguistics*. Kolkata: Sahitya Samsad.

Fillmore, C.J., and B.T.S. Atkins. 2000. Describing Polysemy: The Case of 'Crawl'. In *Polysemy*, ed. Y. Ravin and C. Leacock, 91–110. New York: Oxford University Press Inc.

Gavioli, L. 2004. The Learner as a Researcher: Introducing Corpus Concordancing in the Language Classroom. In *Learning With Corpora*, ed. G. Aston, 31–45. Cambridge: Cambridge University Press.

Ghadessy, M., A. Henry, and R. Roseberry (eds.). 2001. *Small Corpus Studies and ELT: Theory and Practice*. Amsterdam/Philadelphia: John Benjamins.

Granger, S., E. Dagneaux, F. Meunier, and M. Paquot. 2009. International Corpus of Learner English. In *Handbook & CD-ROM. Version 2*. Louvain-la-Neuve: Presses universitaires de Louvain. Available from http://www.i6doc.com.

Granger, S., J. Hung, and S. Peter-Tyson (eds.). 2002. *Computer Learner Corpora, Second Language Acquisition, and Foreign Language Teaching*. Amsterdam: John Benjamins.

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 *Teaching and Language Corpora*, ed. A. Wichmann, S. Fligestone, T. McEnery, and G. Knowles, 100–115. London: Longman.

Kettemann, C.B., and G. Marko (eds.). 2002. *Teaching and Learning by Doing Corpus Analysis*. Amsterdam-Atlanta, GA: Rodopi.

Kirk, J.M. 2002. Teaching Critical Skills in Corpus Linguistics using the BNC. In *Teaching and Learning by Doing Corpus Analysis*, ed. C.B. Kettemann and G. Marko, 183–197. Amsterdam-Atlanta, GA: Rodopi.

Kübler, N. 2002. Linguistic Concerns in Teaching with Language Corpora: Learner Corpora. In *Teaching and Learning by Doing Corpus Analysis*, ed. C.B. Kettemann and G. Marko, 133–145. Amsterdam-Atlanta, GA: Rodopi.

Leech, G., B. Francis, and X. Xu. 1994. The Use of Computer Corpora in the Textual Demonstrability of Gradience in Linguistic Categories. In *Continuity in Linguistic Semantics*, ed. C. Fuchs and B. Vitorri, 31–47. Amsterdam, Philadelphia: John Benjamins.

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 Versus traditional human teaching methods of part of speech analysis. *Computer Assisted Language Learning* 8 (2–3): 259–274.

McEnery, T., and R. Xiao. 2010. What Corpora can Offer in Language Teaching and Learning? In *Handbook of Research in Second Language Teaching and Learning,* vol. 2, ed. Hinkel, 364–380. London: Routledge.

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 *English Corpus Linguistics: Studies in Honour of Jan Svartvik*, ed. K. Aijmer and B. Altenberg, 182–196. London: Longman.
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.
Schütze, H. 1997. *Ambiguity Resolution in Language Learning: Computational and Cognitive Models*. Cambridge: Cambridge University Press.

Web Links

http://mdavies.for.ilstu.edu/sintaxis.
http://www.teachitelt.com/interactives.
https://sites.google.com/site/eltmethodologies/approaches/data-driven-learning.
http://eurocall.webs.upv.es/documentos/newsletter/papers_20(1)/07_boulton.pdf.
http://alannahfitzgerald.org/tag/data-driven-learning-ddl/.
http://ojs.academypublisher.com/index.php/tpls/article/view/tpls020715261531/5070.
http://www.proz.com/translation-articles/articles/633/.

Chapter 7
Corpus as a Secondary Resource for ELT

Abstract In this chapter, we propose for utilizing English Language Corpus (ELC) as a secondary resource for developing English Language Teaching (ELT) materials for teaching English to non-native learners. We argue for using ELC as one of the most authentic representative collections of modern English language from where we can extract necessary linguistic data, appropriate information, and suitable examples to develop some of the basic as well as additional resources of ELT, such as textbooks, reference materials, syllabuses, grammar books, dictionaries, and terminology databases, for the learners. Keeping the requirements of individual learners in mind, we also propose for customizing the materials to make these maximally useful within the broader scheme of computer-assisted language teaching (CALT) method. In our view, the ELT resources developed from the ELC can be far more useful for the learners, because the ELC represents diversified varieties of usage of modern English in real life situations, which the learners need to assimilate to be at par with native speakers across the globe. Also, ELC-based ELT resources can be far more interesting and beneficial to the learners if the ELT instructors carefully utilize these resources keeping in mind the requirements of specific learners. If the ELT learners want to compete in the global frame, then utilization of ELC-based ELT resources is the most useful solution, which we can ignore at the cost of our own peril.

Keywords Corpus · Primary resource · Secondary resource · Dictionary Grammar · ELT learners · Idioms

7.1 Introduction

During last few years, we have noted that the English Language Corpora (ELC), due to their high representational value is used as one of the most reliable resources for English Language Teaching (ELT) at different levels of proficiency of the learners (Biber 1996). In fact, the idea of teaching English without reference to ELC has become a non-reliable proposition, as data, information, and examples obtained from ELC provide authenticity and reliability in the entire process of teaching English.

N. S. Dash and L. Ramamoorthy, *Utility and Application of Language Corpora*,
https://doi.org/10.1007/978-981-13-1801-6_7

This inspires scholars to utilize the ELC in different ways for teaching English to the learners (Botley et al. 2000; Granger et al. 2002). In most cases, the ELC is used for selecting common choices of linguistic usage over those that are less common or less frequent in use in English (Kübler 2002). Also, the ELC is used to revise our traditional ELT text materials as these resources usually contain imaginative information the descriptions of which largely differ from the actual use of English noted in ELC.

In principle, all ELT materials should be maximally empirical and usage based, because this is what the learners need to learn. In this case, ELC becomes highly relevant as it provides a wide variety of attested patterns of empirical usage of language elements the learners need to assimilate. Therefore, ELC is an indispensable resource in designing ELT syllabus, textbooks, and course materials (Barlow 1996). Since ELC-based texts refer to more common and frequent choice of words over rare ones, the learners gather valuable information regarding the texture and structure of patterns of use of words in English (Gavioli 2004). Keeping in view the multiple advantages of ELC in preparation of basic and additional ELT materials for English learners, in this chapter, we discuss about using the ELC as a secondary resource for retrieving data, information, and examples to develop ELT textbooks, bilingual dictionaries, dictionary of idioms, phrases and proverbs, and grammar books for the Indian learners.

In Sect. 7.2, we make a distinction between ELC as a primary resource and ELC as a secondary resource; in Sect. 7.3, we discuss how ELC may be used for developing ELT textbooks; in Sect. 7.4, we propose for developing corpus-based digital bilingual dictionaries between English and Indian languages; in Sect. 7.5, we propose for compiling corpus-based dictionary of idioms, phrases, and proverbs; and finally, in Sect. 7.6, we suggest for developing corpus-based ELT grammar books for English learners. We argue that until and unless these resources are developed from ELC, the teaching of English to the English learners will not be much useful and beneficial.

7.2 Primary Resource Versus Secondary Resource

In our view, the ELC available in digitized and processed form can be highly useful and effective both as a primary and a secondary resource for teaching English to the learners. When we envisage ELC as a primary resource, we visualize learners engaged in extracting relevant linguistic data, examples, information from ELC and using them directly in their classroom learning activities for enhancing their linguistic knowledge and communication skills in English (Johns 1997). In this scheme, teachers and learners can jointly access ELC directly in classroom teaching situations (Dash 2011).

On the other hand, when we envisage ELC as a secondary resource, we visualize ELC as an authentic source of widely varied source of linguistic data, examples, and information wherefrom ELT materials can be produced to be used by learners

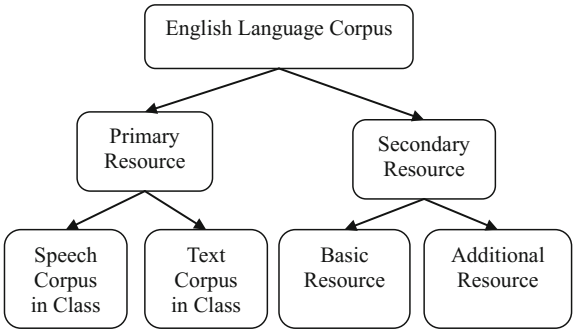

Fig. 7.1 Use of ELC in ELT purposes for English learners

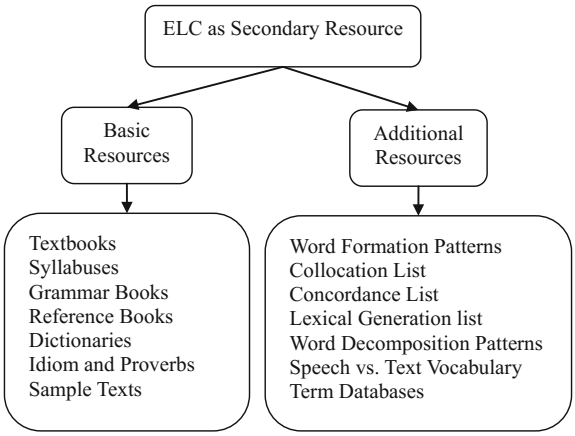

Fig. 7.2 ELC as a secondary resource in ELT

(Granger et al. 2002). In the first scheme, which we have addressed in the earlier chapter (Chap. 6), both teachers and learners can access ELC directly in a classroom situation. In the second scheme, which we address here, ELC is a database to be utilized for developing ELT textbooks, bilingual dictionaries, reference materials, and grammars for learners (Hunston 2002). The two-way utilities of ELC may be visualized from the diagram given below (Fig. 7.1).

When we visualize ELC as a secondary resource, we want to utilize ELC as a huge and varied storehouse of linguistic data, information, and examples from where all kinds of ELT materials may be produced as the 'knowledge-bank' for the learners (Granger et al. 2002). In this scheme, all kinds of ELT textbooks, dictionaries, reference materials, and grammar books are to be developed for the learners (Hunston 2002). Also, it is to be used for developing useful teaching aids, such as word formation patterns, collocation list, concordance list, which may be used by learners as off-classroom reference materials. The two-type utility of ELC as a secondary resource may be visualized from the following diagram (Fig. 7.2).

7.3 ELT Text Books

The ELT textbook developers can design course materials in an effective way by using ELC, because 'language looks different when we look at a lot of it at once' (Sinclair 1991: 100). The textbooks may be called corpus-based textbooks, which are meant to be based on data and information retrieved from ELC (Sinclair and Renouf 1988; Wills 1990). It refers to a method that differs from conventional syllabus in the sense that its central theme is nothing but 'organization of lexis' in a piece of text. In a simple sense, it makes sense to teach the frequently used words found in ELC to learners, since the aim of ELT is to project on common words, their central patterns of usage, and their combination of all types (Sinclair and Renouf 1988: 148). As we do not endorse the traditional practice of acquiring a large vocabulary, especially, at the initial stage of learning, we argue for making full use of common words a learner needs to learn at a particular stage of learning. There is far more general utility in showing the usage and distribution patterns of a few frequently used words than adding new or less known words in the basic vocabulary of learners (Sinclair and Renouf 1988: 155).

Within this frame, our argument is that if learners are taught with textbooks made with frequently used English words along with their patterns of distribution and usage, learners will acquire a better knowledge of modern English. When they learn the texts, they learn the most common usage patterns of the most frequently used English words, and thus, they succeed to cover the central structure of the modern English. For instance, the *British National Corpus* and the *Bank of English* shows that the most frequent use of the English verb 'make' is found in combinations, such as *make noise, make decisions, make discoveries, make profit, make arrangements, make love, make fool, make road, make a choice*. The verb is less frequently used in the sense of constructing or building a thing, such as *make a cake, make a house, or make a doll*. According to Sinclair (1991: 101), in present English texts, the verb 'make' is more often used as a 'delexicalized verb' than as an ordinary lexical verb. An ELT textbook that primarily shows the most frequently used sense of the word actually gives learners an opportunity to learn expressing sophisticate meanings with a simple verb in English.

In the ELC-based textbook, we do not need to depend on separate grammar books, because what is traditionally called a 'grammar' is nothing but the 'patterns of word use' (Wills 1990: 51). In other words, the most productive way of interpreting English grammar to learners is to show them the patterns of use of English words. It again includes only the most frequently used English words in the textbook. Since the patterns of use of common English words are always tagged with supporting examples, learning the words means learning the patterns of their usage, and therefore, learning the grammar of English. In our view, if examples of English words (and phrases) with reference to their patterns of use in ELC are presented before learners, they will necessarily learn all the relevant rules, methods, and principles of English grammar. For instance, the tense system of English verbs, which is considered as one of the main organizing features of an ELT course, may be presented before the learners as

combinations of some common words used in different time frames to enrich their understanding of English grammar (Sinclair and Renouf 1988: 155). 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' are better handled as 'aspects of vocabulary.' For example, the passive forms of BE may be seen as 'BE + Adjective' or 'BE + past participle,' rather than as a transformation of the active forms (Wills 1990: 17). Similarly, the conditional forms may be handled by looking at the hypothetical meaning of 'would,' rather than by proposing a rule about the sequence of tenses, that often fails to work (Wills 1990: 18–19).

In ELC-based ELT, a textbook is prepared with actual citation of English usages as found in the ELC. A text developer uses authentic English texts containing instances of the most frequent patterns of use of common words in English so that he can convincingly refer to these instances to exemplify what the learners need to learn. The job of an ELT teacher is to follow the textbooks and encourage the learners to engage them with English texts to help them 'notice' the patterns of use of English words (and phrases) in the texts. In essence, the description of patterns of words used in ELC is equal to the description of the language. If a textbook contains a list of English words, it is the list of the most frequently used words found in ELC with close reference to their typical idiomatic, collocational, and phrasal usages. Once an ELC-based textbook is taught to the learners, an ELT syllabus is invariably covered.

In our view, this method can redefine the role of textbook designers quite considerably. A textbook designer, instead of selecting specific English texts for reference, description, and illustration, can select interesting text samples from ELC keeping a check on the balance of overall collection of samples to ensure that the most frequent words and their typical usages are covered which the learners require. The textbook designer may consider different genres of ELC to be varied and appropriate with regard to age and need of the learners. This kind of subjectivity will provide an appropriate answer to the objection often raised by critics that ELT textbooks lead to the generation of artificial teaching materials as English texts are written especially to demonstrate the form and function of keywords of English to learners (Long and Crookes 1992: 33). The basic idea of the 'lexical syllabus' as proposed by Wills (1990) relates to a text database that constitutes an authentic collection of texts. This is, indeed, not dissimilar from 'task-based syllabus' proposed in Long and Crookes (1992).

7.4 Bilingual Dictionary

An important aid of ELC-based ELT is a generation of a bilingual dictionary, the lack of which is a real bottleneck in the present context of ELT in India and other countries. The presently available bilingual dictionaries can hardly compensate this deficiency since they do not contain enough information about lexical usage patterns, different categorization and distribution of English words across genres and text types, and conceptual equivalents from the learners' mother tongue. In this case, a

Table 7.1 Sense mapping between English and Bangla words for conceptual equivalents

English	Domain	Sense	Bangla Equivalent
Delivery	Medical science	To give birth a baby	janma deoyā
	Classroom	To teach something	baktṛtā deoyā
	Mass rally	To speak at a public gathering	bhāṣaṇ deoyā
	Courier service	To carry and give something	põuche deoyā
	Cricket	To throw a ball	bal karā

bilingual dictionary compiled with information about the patterns of use of English words extracted from ELC can have tremendous usefulness for learners, as it shows how English words vary in part-of-speech and lexico-semantic function due to their occurrence in different contexts, genres, and text types. We can compile bilingual dictionaries between English and respective target languages (e.g., *English–Bangla*, *English–Hindi*, *English–Tamil*) with conceptually equivalent words as they share many common linguistic and semantic properties. The conceptually equivalent words are those which are equal in sense, content, and implication even if they differ in phonetic representation and orthographic form (Dash 2005a, b, c: 363).

With regard to content, a bilingual dictionary includes a list the most frequently used words of English with their conceptual equivalents collected from the learners' languages (say, *Bangla*, *Hindi*, *Tamil*, for example) keeping in mind their recurrent utilities and relevance in teaching English to the learners. Such a dictionary not only provides cognitive-cum-conceptual equivalents both in source and target languages but also supplies the necessary information for selection of equivalent forms appropriate to particular contexts marked with register and discourse variations. Keeping these utilities in mind, if we think of compiling an English–Bangla bilingual dictionary, the following features will receive utmost attention:

(1) It will include the most frequently used English words collected from the ELC.
(2) Words collected from ELC will be sorted in alphabetical order.
(3) Words should be classified in sense-related frames with citation of their usage in contexts.
(4) Each English word will include at least one conceptually equivalent word from Bangla.

More than one equivalent word from Bangla may be provided if the contextual variation of use of the English words is taken into consideration. For instance, the English word *delivery* may be supplied with at least five Bangla equivalent words, so that the Bangla learners know how the meanings of the English words differ depending on contexts of their occurrence. While learning meanings of the English words, selection of an appropriate equivalent word from Bangla will depend on the domain of occurrence and sense of the English word (Table 7.1).

The list (Table 7.1) reveals that the sense of the word 'delivery' differs depending on the domains of its occurrence in English. What it implies in the domain of medical science is different from that of the domain of education, politics, business, and

Table 7.2 Concordance of word for classification of sense variations

An exiting	game	was played between the two teams
The winner was declared in the last	game	of the match
They won the competition in the last	game	
The next Olympic	games	will be held in London
The annual	games	and sports were held in December
The couple is not new to this	game	of love
The final	game	ended in six-all
They lost but played a good	game	
I was just playing a	game	with you jokingly
It is none of your	games	
So that was your	game	which I failed to understand
The	game	of the authority annoyed the government
It was a wild	game	in last night's party
The hunters went to the	game	reserve in a group at night
The hound chased the	game	into the wild grass
They knew the	game	was over for their leader

sports. This signifies that by considering domain-specific senses of English words, we can compile a bilingual dictionary and select appropriate words from the Indian languages to supply equivalent forms for the English words. This is possible only when we access ELC through concordance to track a possible range of sense variations of the most frequently used words, calculate frequency counts of different senses, and identify the contexts responsible for triggering sense variation. Extraction of information of this kind from ELC can help us compile a database of lexical sense variation the utilization of which can enhance English learning skill of the learners.

The extraction of information of sense variation words from ELC can help us compile a concordance list which can be used for compiling a bilingual dictionary. Since extraction of conceptually equivalent words from ELC is not a complex task, it does not require any special skill in corpus management. We can do it with a workable knowledge of word concordance to retrieve examples in the following manner (Table 7.2).

Search for sense variations of a word normally begins with identifying a particular word occurring in a specific context and expressing a particular sense which differs from its etymological or dictionary sense (Sinclair 1991). After all senses of a word are collected from a corpus through a concordance, the work for classifying senses and identifying contexts that trigger sense variations start (McEnery and Wilson 2001). It involves identification of the total senses and the most frequent sense. When a corpus yields different senses as alternative candidates, we have to make a choice among equivalent senses on the basis of their frequency of occurrence to

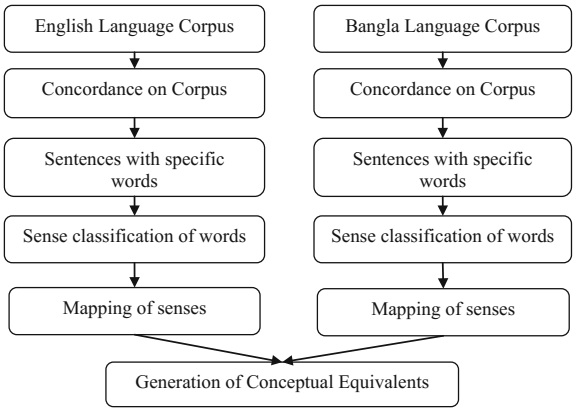

Fig. 7.3 Generation of conceptual equivalents from corpora

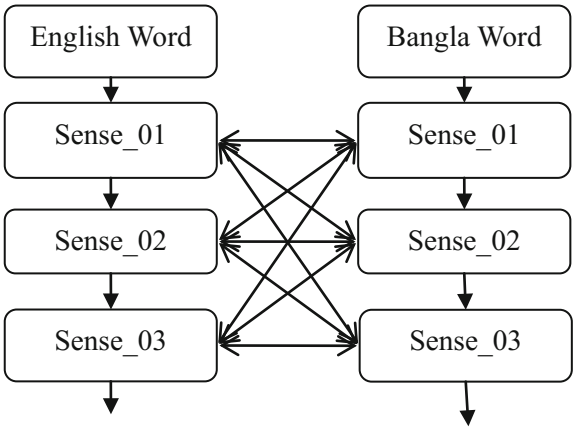

Fig. 7.4 Sense mapping between the words in two languages

be included in the bilingual dictionary. Finally, classification of senses in different categories provides necessary insights to understand how senses are interlinked and should be placed in the dictionary (Fig. 7.3).

Mapping of senses is a different task that supplies necessary directions to assemble conceptually equivalent words or word pairs from Indian languages to be included in the dictionary. The analysis of senses of words found in the concordance list will lead us to develop a dictionary of conceptually equivalent words from the languages (Fig. 7.4).

It is found that equivalent words seldom mean the same thing in all contexts. Even within closely linked languages, the so-called 'equivalents' are seldom equivalent with regard to their distribution and senses (Dash 2005b). Semantic equivalence is possible with the names of animals, tools and scientific terms, but hardly with ordinary words (Landau 2001: 319). This signifies that selection of equivalent words from learner's language for ordinary English words requires a high level of linguistic expertise on the part of a bilingual dictionary compiler for generating useful outputs.

Therefore, extraction of words from ELC and providing equivalent words from learner's language requires assessment and validation by expert linguists. The role of a bilingual dictionary maker is to collect frequently used words from ELC, collect equivalent words from respective learners' languages, classify the senses based on their usages in different contexts, map senses of English words with that of words of learners' languages, and compile a sense-paired bilingual dictionary.

The basic pedagogical purpose of a bilingual dictionary is to supply learners the required information about forms and functions of English words they need to acquire for learning the language. They do not need prior lexical knowledge of English to start using a bilingual dictionary as a supporting tool. Although this issue has not been raised by ELT teachers engaged in teaching English to the learners, corpus processing tools can help them to develop this resource for the learners. An English–Bangla bilingual dictionary, for example, may be replaced with a bilingual thesaurus, which can record a wide range of English text types to provide access to the contextual frames of words used in corpora by alphabetical and thematic indexes of lexical items of English.

The role of an ELT teacher, in this frame, is to present the resource to the learners. He has to monitor that the dictionary does not give excessive lexical load on the learners. Thus, an ELT teacher, with the help of a bilingual dictionary can lead learners to proceed toward a more advanced stage of learning where learners themselves can explore more intricate patterns of use and sense variations of English words.

7.5 Bilingual Dictionary of Idioms and Proverbs

The familiarity with the wide range of English idioms and proverbs and the ability to use these forms in an appropriate manner are some of the distinguishing marks of native-speaker-like command over English, which all the learners want to have. Scholars always emphasize on the importance of learning English idiomatic and proverbial expressions as a necessary means in learning good English: '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, and insipid' (Smith 1925: 276). 'Proverbs continues—the early collectors never tired of stating—to provide the sauce to relish the meat of ordinary speech' (Simpson 1982: x–xi).

Although many people in India and other non-English countries can speak and write English correctly at the grammar and syntax level, most of them can hardly express themselves 'idiomatically' in a way the native English speakers spontaneously pick up and choose idiomatic and proverbial expressions in their written and spoken texts. Consequently, what Indians and other non-English speakers speak or write often sounds bit awkward, flat, and lengthy. For example, while Indian speakers usually find it easier to use non-idiomatic words, the native English speakers often tend to use idiomatic expressions, phrasal verbs, and proverbial expressions quite easily (Xiao-Jun 2003: 296). Full of vitality, these idiomatic and proverbial

expressions tend to fill in the blanks in actual communication and thus make it easier for the people to exchange thoughts because native people use them without the special effort of the formulation. Thus, it becomes clear that learning English idioms and proverbs (especially those with figurative senses) should be treated as an essential part learning English, because 'learners find these puzzling idioms their main difficulty in learning English' (Henderson 1954: 5). In our view, it is not possible to acquire a thorough knowledge of English without being familiar with the idiomatic and proverbial expressions as well as with slang and vulgarism used in it.

The English idiomatic expressions, set phrases, and proverbial expressions can be best obtained from ELC tagged corpora with detailed information of their contextual usages. A well-developed database of this kind can help learners understand the actual figurative senses of these expressions as well as to learn patterns of use of these forms within natural communication environments. If we can compile a bilingual database of English idioms and proverbs and map it with equivalent expressions from Indian languages, it will tremendously help Indian learners as it will carry authentic context-based senses expressed in English idioms, phrasal, and proverbial expressions. Conceptually equivalent forms provided from Indian languages will help learners to access and understand English expressions in a sensible manner (Dash 2007). An ELT teacher may, if he wants, integrate domain-specific information and usage patterns of these expressions in the database for better comprehensibility.

For compiling a database of English idioms, phrases, and proverbs from ELC, we need to use simple data retrieval techniques applied on language corpora. While applying this method on ELC, our basic goals are to do the following tasks:

(1) Collect the most frequent English idioms, phrases, and proverbs from ELC.
(2) Decipher actual senses of these forms expressed in the contexts of their usage.
(3) Classify the senses based on an appropriate scheme of sense classification.
(4) Find out the appropriate matches of these from Indian language corpora.
(5) Supply appropriate matches for generating a bilingual dictionary of idioms and proverbs.

We do not expect hundred percent success in finding out conceptually equivalent expressions from the Indian languages for English idioms, phrases, and proverbs because these languages are different in many aspects, features, and properties. However, minute scrutiny of Indian language corpora (Dash 2001) may help us collect a large number of expressions, which appear to be conceptually equivalent or nearly equivalents to their English counterparts, as the following English–Bangla idioms and proverb pairs show (Table 7.3).

In essence, we need to compile a dictionary containing frequently used English idioms, phrases, and proverbs from ELC with their conceptually equivalent forms obtained from Indian languages. Since such an attempt is never made before to present English multiword expressions with their equivalents from the Indian languages with direct reference to corpora, either in electronic or printed form, we advocate for creating these resources keeping in mind the need of advanced learners who want to acquire mastery over English with all its intricacies and semantic nuances. We believe that until and unless the learners are able to understand the actual meanings hidden

Table 7.3 English idioms and proverbs and their Bangla equivalents

English idioms/proverbs	Bangla equivalents
Apple of one's eye	cokher mani
Shedding crocodile's tear	kumbhirāśru barjan karā
To make a bedlam	narak guljār karā
To bell the cat	birāler galāy ghaṇṭā bāsel1dhā
Blueblood	nīl rakta
Bolt from the blue	binā meghe bajrapāt
To paddle one's own canoe	nijer carkāy tel deoyā
On cloud nine	sukher saptam svarge
A cock and bull story	āṣāre galpa
A white elephant	śvet hastī
By hook or by crook	ýena tena prakāreṇa
Horns of a dilemma	ubhay saṅkaṭ
To add insult to injury	kāṭā ghāye nuner chiṭe deoyā
To carry coal to New Castle	telā māthāy tel deoyā
Too many cooks spoil the broth	adhik sanyāsīite gājan naṣṭa
Once in a blue moon	kāle bhadre
In the nick of time	śeṣ samaye
Pour oil on troubled water	agnite ghṛtāhuti deoyā
Raining in cats and dogs	muṣaldhāre bṛṣṭipāt
Black sheep of the family	baṃśer kulāṅgār
Writing on the wall	deoyāl likhan
To cry in wilderness	araṇye rodan karā

under surface forms of English idioms, phrases, and proverbs, their basic knowledge of English is not adequate and complete. In such a situation, this dictionary will invariably enrich learners with knowledge about the modern English language.

In the context when we find that the majority of ELT materials used for the Indian learners contain a large number of English idiomatic, phrasal as well as proverbial expressions used in different senses, a dictionary of this kind can be used as an excellent resource for the Indian learners. We can also use this for the purpose of translating English text into the Indian languages in manual and machine translation where a dictionary of conceptually equivalent idioms, phrases, and proverbs of both the languages is indispensable for faithful translation outputs (Dash 2004). After completion of the database, we propose to rearrange the entire database in reverse order where Indian languages will be used as the source language and English as the target language. In this case, these expressions may be rendered in the *International Phonetic Alphabet* (IPA) so that people who are learning Indian languages as a second language are able to pronounce the words correctly. This venture may be started after the completion of the first phase.

Supplying authentic intralinguistic and extralinguistic information of multiword expressions like idioms, phrases, and proverbs in classroom situation may ask for an online interactive interface for the bilingual dictionary. To meet this demand, we need to develop the dictionary in an electronic form and upload it on the internet in such way that a simple hit on a particular expression can produce a string of equivalent forms stored in the dictionary for the learners. Moreover, it should produce all word-related information and examples (e.g., orthography, spelling, pronunciation citation, usage, illustration) of the expressions both from English and Indian languages. If we can build up this resource, teaching English will be more interesting and learners will greatly benefit by exploiting the dictionary. The role an ELT teacher will be that of a 'class coordinator' who will be much relieved from the load of teaching he usually carries out in traditional classroom teaching situations.

7.6 ELT Grammar

The ELC is a highly useful resource for providing morphological and syntactic information of various types necessary for writing ELT grammars (Halliday 1991). Such grammars can be used effectively for the Indian learners since these can provide information about various grammatical issues observed in modern English as well as can supply empirical data for testing earlier hypotheses about various grammatical rules of English. Normally, an ELT grammar book is based on the intuition of a grammarian about the language rather than on information and examples of actual English use. Therefore, whatever observations a grammarian makes and however fantastic these observations appear, these are not beyond verification with evidence of actual performance of native English users. Even the generative grammarians are not willing to agree with assumptions of intuitive grammarians if these assumptions are not verified with examples of real empirical texts.

The availability of ELC like the *British National Corpus*, the *Bank of English*, the *American National Corpus* has made it possible for us to record performance-based grammatical features of English to be presented in ELT grammar books. It has been strongly argued that '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). To substantiate the relevance of corpus-based grammars in ELT for non-native speakers, a highly acclaimed English grammar is already developed from ELC within last few years which can also be used for this purpose (Quirk et al. 1985).

Following this trend, we suggest for using ELC for developing ELT grammars for the Indian learners. As these corpus-based grammars differ from traditional grammars not only in the proposition of theories but also in the formation of rules and principles, they faithfully reflect on the English language as it is actually used by native speakers. Thus, a corpus-based grammar, if used intelligently, can become highly useful for the Indian learners. Moreover, since ELC contains samples from imaginative to informative texts covering almost all the domains of English use, a grammar book

developed with data, information, and examples from ELC can have wider coverage than a traditional intuitive grammar book can hope for.

According to our scheme, in a corpus-based grammar, each lesson will be supported by data and information obtained from ELC. This gives the learners a great opportunity to learn how the use of English sentences varies based on domains and genres of discourse. They will also learn how variation in use of words, idioms, phrases, and sentences is controlled by the demand of particular discourse. The Indian learners can use such a grammar book to learn mutual as well as exclusive distributions of sentence types across the genres and text types to understand the syntactic varieties practiced in modern English. Thus, utilization of a corpus-based English grammar can enrich the Indian learners to know what kinds of forms of English sentences are used in English based on the type of a text and what kinds of grammatical rules are actually invoked for generating sentences of different types. In essence, a corpus-based English grammar book can be an excellent resource for beginners and advanced learners who aspire to achieve a 'native-speaker-like-command' over the English language—a coveted dream of many generations of India.

7.7 Conclusion

We have made a humble attempt here to present a simple proposition. We have tried to show how ELC can be used as a secondary resource in the act of teaching English to the Indian learners. In a systematic manner, we have tried to highlight how ELC may be used as a resource of linguistic data and information to develop language teaching aids like textbooks, bilingual dictionary, a dictionary of idioms and proverbs, and ELT grammar books. Although many other ELT resources are possible to generate from ELC, these are not addressed here due to the limitation of space.

It is a fact that the introduction of ELC in ELT has made a tremendous breakthrough by lending heavily toward the modification of traditional ELT texts and study materials for the non-native English learners (McEnery and Xiao 2010). However, the unfortunate thing is that, in comparison to other countries, the Indian scholars lag far behind in developing corpus-based teaching resources and utilizing them for teaching English to the Indian learners.

There is no doubt about the application relevance of ELC in ELT. We should, therefore, divert our attention toward this new approach of ELT to rejuvenate the field with a new lease of information. The Indian learners will not only benefit from this approach but also excel in English when they are put into competition with speakers of other countries. However, before we begin to work in this direction, we have to procure authentic ELC of the British English and the American English as well as process them in appropriate manners to be used in ELT resource generation.

References

Aarts, J. 1991. Intuition-Based and Observation-Based Grammars. In *English Corpus Linguistics: Studies in Honour of Jan Svartvik*, ed. K. Aijmer and B. Altenberg, 44–62. London: Longman.

Barlow, M. 1996. Corpora for Theory and Practice. *International Journal of Corpus Linguistics.* 1 (1): 1–38.

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. McEnery, and A. Wilson (eds.). 2000. *Multilingual Corpora in Teaching and Research*. Amsterdam-Atlanta, G.A.: Rodopi.

Dash, N.S. 2001. *Corpus-Based Computational Analysis of the Bengali Language.* Unpublished Doctoral dissertation, University of Calcutta, Kolkata.

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. 2005a. Corpus-Based Machine Translation Across Indian Languages: From Theory to Practice. *Language In India* 5 (7): 12–35.

Dash, N.S. 2005b. Role of Context in Word Sense Disambiguation. *Indian Linguistics* 66 (1–4): 159–175.

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

Dash, N.S. 2007. Indian Scenario in Language Corpus Generation. In *Rainbow of Linguistics*, vol. I, ed. N.S. Dash, P. Dasgupta, and P. Sarkar, 129–162. Kolkata: T. Media Publication.

Dash, N.S. 2011. Use of English Corpora as a Primary Resource to Teach English to the Bangla Learners. *Indian Journal of Applied Linguistics* 37 (1): 7–18.

Gavioli, L. 2004. The Learner as a Researcher: Introducing Corpus Concordancing in the Language Classroom. In *Learning with Corpora*, ed. G. Aston, 31–45. Cambridge: Cambridge University Press.

Granger, S., J. Hung, and S.P. 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 *English Corpus Linguistics: Studies in Honour of Jan Svartvik*, ed. K. Aijmer and B. Altenberg, 24–43. London: Longman.

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 CALL Program. In *Teaching and Language Corpora*, ed. A. Wichmann, S. Fligelstone, T. McEnery, and G. Knowles, 100–115. Harlow: Addison Wesley Longman.

Kübler, N. 2002. Linguistic Concerns in Teaching with Language Corpora: Learner Corpora. In *Teaching and Learning by Doing Corpus Analysis*, ed. C.B. Kettemann and G. Marko, 133–145. Amsterdam-Atlanta, G.A: Rodopi.

Landau, S.I. 2001. *Dictionaries: The Art and Craft of Lexicography*, 2nd ed. Cambridge: Cambridge University Press.

Long, M.H., and G. Crookes. 1992. Three Approaches to Task-Based Syllabus Design. *TESOL Quarterly* 26 (1): 27–56.

McEnery, T., and A. Wilson. 2001. *Corpus Linguistics*, 2nd ed. Edinburgh: Edinburgh University Press.

McEnery, T., and Xiao, R. 2010. What Corpora can Offer in Language Teaching and Learning. In *Handbook of Research in Second Language Teaching and Learning.* vol. 2, ed. Hinkel. E., 364-380. Routledge: London, New York.

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

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 *Vocabulary and Language Teaching*, ed. R. Carter and M. McCarthy, 140–160. London: Longman.
Smith, L.P. 1925. *Words and Idioms*. London: Constable.
Wills, J.D. 1990. *The Lexical Syllabus: A New Approach to Language Teaching*. London: Harper Collins.
Xiao-Jun, H. 2003. Lexicographical Treatment of Idioms and Proverbs. In *Lexicography: Critical Concepts*, vol. II, ed Hartmann, R.R.K. 295–312. London: Routledge.

Web Links

http://eltj.oxfordjournals.org/content/65/1/92.extract.
http://users.utu.fi/micnel/thesis.html.
www.eisu2.bham.ac.uk/johnstf/timeap3.htm.
http://tesl-ej.org/ej32/a1.html.
http://www.lexically.net/wordsmith/corpus_linguistics_links/.
http://iteslj.org/Articles/Krieger-Corpus.html.
http://www.eisu2.bham.ac.uk/johnstf/timeap3.htm.

Chapter 8
Corpus and Dictionary Making

Abstract Recent works show that a dictionary can be made to a certain level of satisfaction if it is made with data and information acquired from widely representative and properly balanced language corpus. A language corpus provides an empirical basis in the selection of words and other lexical items as well as in supplying the most authentic information relating to pronunciation, usage, grammar, meaning, illustration, and other information with which all the words and lexical items in a general reference dictionary are furnished with. In the same manner, a language corpus supplies the most authentic information relating to compounds, idioms, phrases, and proverbs, etc., which is also included within a general reference dictionary with equal attention and importance. In this chapter, we try to explain how linguistic data and information collected from a corpus can contribute toward compilation a more useful dictionary. Although a corpus has better functional utilities in development of electronic dictionary, we like to concentrate here on the use of a corpus in the compilation of printed dictionary. We shall occasionally refer to the TDIL corpora developed in the Indian languages and use linguistic data and information from these to substantiate our arguments and observations.

Keywords Dictionary making · Lexicographic data · Word collection · Corpus
Selection of lexical stock · Headwords · Spelling variation · Part-of-speech
Grammatical information · Definition · Description · Semantic information · Usage

8.1 Introduction

The experiences gathered from the recent works of dictionary making show that it is done better if the dictionary is made with data and information acquired from widely representative and properly balanced language corpora. Corpora accessed with versatile text processing techniques provide a solid empirical base for the lexical items included in a dictionary as linguistic information relating to single words, compounds, idioms, set phrases, proverbs, etc., are best available from corpora. In fact, recent success in corpus-based dictionary leads lexicographers to direct attention

N. S. Dash and L. Ramamoorthy, *Utility and Application of Language Corpora*,
https://doi.org/10.1007/978-981-13-1801-6_8

toward the ordinary use of language (Landau 2001: 278) as it is exemplified that the ordinary use of words, by way of providing lexicological information, supplements the intuition of dictionary makers (Atkins and Levin 1995). Evidence collected from corpora reveals that for many common words, the most frequent meaning is not the one that first takes place in dictionaries (Sinclair 1991: 39). These arguments challenge our traditional methods of dictionary compilation and description of words as well as advocate for obtaining lexical items and their information from corpora to contribute toward the overall growth and maturity of lexicography (Landau 2001: 132).

For developing a good reference dictionary for a natural language, we can have an adequate amount of linguistic data and information from both diachronic and synchronic corpora. We can call up all the usage variations of lexical items to compile a dictionary as well as revise an old one. Also, we can provide a definition of lexical items in a more complete and precise manner as we get access to examine a larger number of examples of usage of words in corpora. The entire work can become far more useful and impressive if we can avail information from monitor corpora (i.e., corpora that grow with time) as they allow us to find out new words that enter into a language as well as identify those words that are changed in usage and meaning over time. Also, we can extract authentic information about the frequency of use of lexical items from monitor corpora to determine the number of lexical items to be included in a dictionary as well as to identify the usage variations of the lexical items with regard to genres and text types—an important piece of information indispensable in dictionary making.

We know that corpora contain a large amount of extralinguistic information with regard to demography, society, culture, history, ecology, heritage, and other aspects of life. Therefore, from such a resource we can get valuable insights to tie up usage of words, idioms, phrases, and proverbs as being typical to the particular region, variety, genre, ethnic group, profession, or age. Moreover, we can refer to corpora to call up various word combinations to establish the relationship between the co-occurring lexical items to explore their underlying semantic interfaces. Thus, we can benefit from corpora to understand the usage of phrases, idioms, and collocations more systematically as well as to provide the same in the dictionary with adequate information for retrieving their specific senses.

A dictionary of a living language can never be complete. The perennial changes in life lead to constant modification of language and up-gradation of lexicological information to be included in a dictionary. Diachronically, it is observed that while some words become old and obsolete, many new words are coined to fulfill the needs of the time. On the other hand, those words, which survive the challenge of time, undergo changes in their forms, meanings, and usages to adapt to new concepts and ideas of life. A dictionary in this transition tries to record death, birth, and metamorphosis of words of a language. Thus, it performs two basic tasks:

(1) It restores the vanishing track of lexical loss through diachronic search for lexical evolution in a language and

(2) It represents the contemporary scenario of lexical usage through synchronic projection into the behaviors of lexical items.

Keeping various formational and functional advantages in mind, we discuss in this chapter about the process of compiling a general reference dictionary (GRD) with data, information, and examples collected from a general corpus. Since such a dictionary is never made in any of the Indian languages, we hope this method can start a new trend in dictionary making in India and inspire dictionary makers to adopt corpus-based approach giving up the traditional methods of dictionary making used for ages. The process of compiling corpus-based electronic dictionary is a more complex method which may be discussed separately later. However, this issue will be hinted in this chapter whenever it is required.

In Sect. 8.2, we discuss the benefits corpus in dictionary making; in Sect. 8.3, we discuss the methods of data collection from corpus; in Sect. 8.4, we discuss the process of selection of lexical stock; in Sect. 8.5, we discuss the process of headword selection, and Sect. 8.6, we address the issue of determining spelling of words which is a tough challenge in many Indian languages; in Sect. 8.7, we address the problem of part-of-speech determination of words with reference to their usage in corpus; in Sect. 8.8, we look into the nature of grammatical information gathered from corpus and presented in dictionary; in Sect. 8.9, we discuss the issues of definition and description of words in dictionary; in Sect. 8.10, we discuss the load of semantic information to be presented with words in a dictionary; in Sect. 8.11, we focus on how examples of usage of words can be collected from corpus and presented in dictionary; and in Sect. 8.12, we discuss how the entire process may be realized in a systematic manner in a language.

8.2 Benefit of Corpora in Dictionary Making

For centuries, a close functional relationship is maintained between linguistics and lexicography. The emergence of the corpus has added a new shade to it. At present, due to mutual cognitive dependency and information sharing, lexicographers are becoming corpus linguists and corpus linguists are becoming lexicographers as the distinction between the two is gradually fading. Lexicographers are taking up corpus as a source of quantitative and qualitative information for the betterment of their own field. On the other hand, corpus linguists are expanding their area to lexicography since dictionary making of any kind is gradually tilting toward empirical evidence gathered from the corpus (Moon 1998; Atkins 1998; Atkins and Zampoli 1998).

In general, a digital corpus can take place of traditional citation files to provide empirical evidence for meaning, use, and other linguistic information of lexical items by way of showing their usage in various contextual environments. Since a corpus due to its ‘cosmopolitan composition’ is able to supply a much greater variety of contexts of lexical items than a citation file can ever think of, a corpus is considered

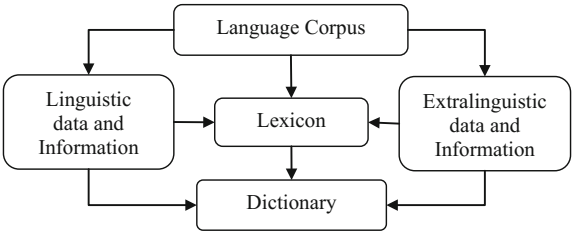

Fig. 8.1 Utilization of corpus data and information in lexicography

the far better source for understanding the lexicological identity of linguistic items and for constructing their more reliable and authentic definitions.

However, there is a problem with regard to a digital corpus. It is the problem of the excessive amount of linguistic data and information. As a dictionary maker, we are burdened with so many examples of the use of a single lexical unit that we are puzzled to select the most appropriate one out of many. In this case, we select those examples, which we consider appropriate for revealing all kinds of lexicological aspects of a lexical unit. This problem is hardly faced with a citation slip as a citation slip usually enlists a few examples of usage of a word. According to some scholars '…the single most striking thing about corpus evidence brought up on a diet of citation slips is the inescapability of the information it presents. If you are confronted with not two or three but dozens or even hundreds of instances of a word being used in a particular way, there is really no arguing with what the corpus is telling you' (Rundell and Stock 1992).

The multipurpose utility of corpus in dictionary making may be represented in the following manner to show how linguistic data and information retrieved from a corpus may help us to accomplish our goal in a far more realistic manner (Fig. 8.1).

The interface between language corpus and dictionary makers in one hand and the dictionary makers and the digital dictionary, on the other hand, may be understood from the diagram given below (Fig. 8.2).

A corpus can be useful for us in several ways in dictionary making. We can sum up how a corpus can help us in the work of dictionary making:

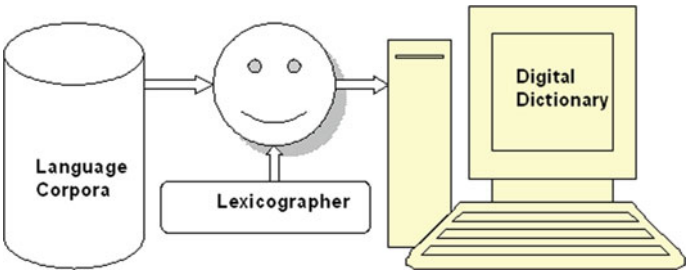

Fig. 8.2 Interface between corpus, dictionary makers, and digital dictionary

(1) Since a corpus is available in machine-readable form, we can extract all relevant, authentic, and typical lexical items with examples of their use from the corpus within a short span of time.
(2) We can extract faithful frequency information of words from a corpus to attest relevance of their entry in the dictionary.
(3) We can retrieve quantified information of collocation from a corpus to show the patterns of lexical combination in a sentence. Information of this sort is useful for text writers and language learners alike.
(4) We can use POS and sense-tagged corpus for word sense disambiguation as well as for more sensible grouping of polysemous and homographic words in the dictionary.
(5) We can retrieve textual (e.g., register, genre, and domain) and sociolinguistic (e.g., user, gender, age, etc.) metadata from annotated corpus to supply an accurate description of the usage of the lexical items in the dictionary.
(6) We can use a monitor corpus to track subtle changes in the meaning and usage of lexical items to keep our dictionary up-to-date.
(7) We can use evidence from a corpus to complement or refute intuitions of early dictionary makers so that the entries in the dictionary are more accurate and authentic.

The above observations are made with the views of Hunston (2002: 96) who argues that corpus has emphasized on five vital issues of dictionary making: emphasis on frequency, emphasis on collocation and phraseology, emphasis on variation, emphasis on lexis in grammar, and emphasis on authenticity (McEnery and Xiao 2010: 266).

8.3 Data Collection from Corpus

For the time being, we can collect linguistic data from a text corpus since a speech corpus in many languages (including those of India) is not yet made. A large, balanced, and adequately representative text corpus contains samples from all possible sources of language use, such as creative writings, contemporary literature, natural science, social science, technology, arts, humanities, business and commerce, medicine, newspapers, magazines, periodicals, posters, advertisements, reports, court proceedings, personal diaries, personal letters, official letters, and similar other resources. Lexical items extracted from a corpus carry detailed contextual information of their occurrence in texts. This is required for expressing the meaning of the lexical items clearly and unambiguously in the dictionary with full reference to their contexts of use. There are two qualities of a good context: 'brevity' and 'clarity.' While 'brevity' is related to the problem of space within a dictionary, 'clarity' is linked with the idea of explicitness of meaning and sense of words.

While collecting words from a text corpus, we should keep in mind to include lexical items relating to all possible situations and domains. Thus, the primary word list

covering various semantic domains and grammatical classes should include words, terms, idioms, phrases, proverbs, collocations, and other lexical items relating to life and living of the people.

Obviously, such an elaborate list will contain a large number of lexical items, which are unknown, not in frequent use, or maybe less comprehensible to the users (Boguraev and Pustejvsky 1996). Also, some of the items collected from a corpus may appear artificial or foreign to a native speaker. Both artificiality and foreign flavor of items may be verified with another corpus. In reality, the collection of lexical items from a corpus confirms their existence in life, language, and culture of the people whose language is represented in the corpus.

This list, however, is not an exhaustive one. It is rather a workable list, which may be compared, verified and augmented with other lexical lists available in the language. This collection of lexical units from different sources will provide greater possibilities for wider representation of a language. If required, the database may be further supplemented with data available from spoken discourses such as narration, eyewitness accounts, instructions, reminiscences, conversations, arguments, dialogues, mimicries, and onomatopoeias (Samarin 1967: 208).

8.4 Selection of Lexical Stock

Selection of lexical stock for a dictionary is a useful but crucial task. We have to select those lexical items that we should include in the dictionary. Also, we have to identify the lexical items that we should ignore. We generally depend on the feature of the frequency of lexical items in a corpus on a logic like this: The more frequent one is an automatic choice while the less frequent one has a marginal chance. In support of this observation, Landau (2001) makes the following comment that seems to be pertinent:

> If one has a 100-million word corpus that is based on a wide variety of written and spoken sources of many different types, and it does not include a single example of a particular lexical item, this datum means something. It does not mean that the lexical item does not exist. No corpus can prove that a word or expression does not exist, and no corpus is perfect. But if the corpus has been put together carefully to be representative, one can conclude that the lexical item, if it exists, either is extremely uncommon or is used almost exclusively in a specialized field that the corpus does not cover. A low frequency for a lexical item is therefore sound justification for omitting it from one's dictionary. Conversely, a high frequency argues strongly for inclusion. Clearly, making such determinations depends upon one's faith in the representativeness of the corpus, a critically important concept in corpus lexicography. (Landau 2001: 297)

While using a corpus for dictionary making, we are aware of the fact that many lexical items, idioms, phrases, and fixed expressions occur rarely in a language. Therefore, we need a very large and widely representative corpus so that we can find citations of the rarely used linguistic items to include in our database. Also, we have experienced that many idioms and fixed expressions that we intuitively feel to be

quite common in use are, in fact, very rare in occurrence. For instance, the English idiom 'kick the bucket' is often assumed to be quite frequently in use. However, analysis of corpus reveals that it occurs mainly in textbook discussions of idioms and phrases and hardly occurs in genuine speech and writing. It has occurred just for 13 times in 100-million *British National Corpus* (Grant 2003: 173). This implies that although some lexical items are easily found in dictionaries, they hardly occur in the text of regular language use.

In general, while collecting lexical stock from a corpus to be used in a general reference dictionary, we can collect following types of words from the corpus: common words, new words, obsolete words, archaic words, scientific and technical words, proper names, empty words, function words, compound words, reduplicated forms, idioms, set phrases, proverbs, quotations, clichés, acronyms, abbreviations, colloquial forms, slang, jargons, cants, affixes, and other lexical items. All these types of lexical items are suitable candidates to be considered to be included in a general reference dictionary.

The data and information that are available from a text corpus for a general reference dictionary include headwords, pronunciation, spelling, part-of-speech, morphological information (e.g., sandhi, assimilation, voicing, compounding, reduplication, etc.), grammatical information (i.e., derivation, inflection, affixes, etc.), definition, synonyms, polysemy, lexical generativity, collocation, usage (general use, idiomatic use, phrasal use, and proverbial use), illustration, and citation (Landau 2001: 278). It is not necessary that each headword should have information all the fields stated here. Some may have synonyms while others have the only illustration. But in most cases, information related to most of the fields is indispensable for a general reference dictionary.

8.5 Headwords

A headword is a 'citation form' or an 'entry word.' The form and meaning are the main criteria for selection of a headword. It is a canonical form which is derived by lemmatization of inflected forms. We can use a corpus to extract three types of lexical items to be used as headwords in the dictionary:

(a) **Single-word units**: It includes both inflected and non-inflected forms. We can put within this list all common words, stems, bases, roots, modified words, sandhi-made words, new words, old words, archaic words, obsolete words, rare words, dialect words, folk words, slang, jargon, cants, taboo words, codes, scientific and technical terms, foreign words, native words, local words, colloquial words, portmanteau words, analogically formed words, acrostic words, proper names, person names, place names, items names, object names, function words, clichés, postpositions, etc.

(b) **Multiword units**: This also includes both inflected and non-inflected lexical items with space between the members that constitute the final surface form.

In this list, we can put several compound words, reduplicated words, idioms, phrases, proverbs, collocations, quotations, set expressions, other longer forms, etc.

(c) **Nonword units**: This normally includes those **lexicographic words** which are not considered as regular words but treated as word-formative elements. In this list, we include empty words, abbreviated forms, acronyms, echo words, clipped words, affixes, inflections, person markers, aspect markers, case markers, plural markers, tense markers, articles, enclitics, particles, acrostic words, and similar other word-formative elements. Although these are not words in the true sense of the term, these are unique lexical items having specific linguistic entities and functions in a natural language.

The single and multiword units are available as free forms, and nonword units are available as bound forms. Some words may be available in new meaning or new part-of-speech. So there is no chance for eliminating any of the items since all these are of equal importance for a general reference dictionary. However, before these are collected from corpus they should be lexically and grammatically tagged, else much information relating to their part-of-speech and meaning will be lost if these are isolated from the context of their sentential occurrence. Such loss is unbearable since we shall lose valuable insights into the functional nature and behavior of the items while they occur in texts.

We can collect many tokens of inflected and non-inflected forms of words from a corpus. While we can directly process non-inflected tokens for obtaining types with a sticker on their total occurrence in each type, we can lemmatize inflected forms to separate affixes and inflections from roots and stems. Furthermore, we can tag all lemmas with stickers of their total occurrence in the corpus, so that information of their frequency of occurrence in the corpus is safely preserved. In the same manner, we can retrieve frequency information of isolated word-formative elements (e.g., case markers, particles, articles, enclitics, etc.) from the total number of tokens for reference in the dictionary. We can use the canonical forms as headwords and non-canonical forms as subentry or run-on words. For instance, while the canonical form *din* 'day' is a headword, forms like *dinkāl, dinrāt, dinagata, dinakṣay, dinrātri, dinānta* may be used as subentry words.

8.6 Spelling Variations

Spelling variation is an important factor that cannot be ignored in case of dictionary development for the Indian languages. In case of languages like English and German, the problem of spelling variation of headwords is not a great problem, as most of the words have one accepted spelling. Obviously, there are certain words in a language like English where some words show two or more spellings. For instance, *color: color, night: nite, light: light, meter: meter, center: center, behavior: behavior, localize: localize*, etc. In such cases, it is known that the first set is used in the British English

Table 8.1 Example of some Bangla words with spelling variation

Words with spelling variations	Gloss
rāni, rānī, rāṇi, rāṇī	Queen
cāprāś, cāprāśi, cāprāśī, cāprās, cāprāsi, cāprāsī	Servant
beṅāci, beṅgāci, byāṅāci, byāṅgāci, beṅāchi, beṅgāchi, byāṅāchi, byāṅgāchi	Tadpole
ghumana, ghumano, ghumāna, ghumāno, ghumona, ghumono, ghumuna, ghumuno	Sleeping
beyādapi, beyādapī, beyādabi, beyādabī, be-ādapi, be-ādapī, be-ādabi, be-ādabī, beādapi, beādapī, beādabi, beādabī	Obstinacy
kṛścān, kṛscān, kṛṣṭān, kṛṣṭān, kriścān, kriscān, kriṣṭān, kriṣṭān, khṛṣṭān, khṛṣṭān, khriṣṭān, khrīṣṭān, khriṣṭān, khrīṣṭān	Christian
kalikātā, kalkātā, kolkātā, kyālkāṭā	Kolkata
bāṅlā, bāṅglā, bāṅgālā, and bāṃlā	Bengali

while the second set is used in the American English. The dictionary makers of these two respective varieties of English have a little problem to deal with spelling. Once they decide in which variety they are going to develop a dictionary, they opt for spelling accepted in that particular variety.

But in case of the Indian languages like Bangla, Tamil, Telugu, Malayalam, Santali, and others the phenomenon of spelling variation of a canonical form is a serious problem for the dictionary makers. There are large numbers of lemma which exhibit multiple spellings most of which are accepted and used in standard practice. For example, we can refer to some Bangla words with spelling variations that are collected from a Bangla text corpus (Table 8.1).

It is a real tough job to select a particular spelling out of multiple variations because the selection of any one particular spelling out many variations will raise a question regarding the preference of one against the other. In such a situation, a text corpus can play a crucial role. Information about the relative frequency of variants can lead us to decide which spelling we should select. The best option is to put a corpus in frequency calculation of spelling variation of each word. A particular spelling, which has the highest use in the corpus, may be selected. For instance, in case of Bangla, we may select the spelling *hala* 'was/happened' as the suitable candidate over other variants (*halo, hola, holo, ha'la,* etc.), since this particular spelling records the highest occurrence in all text types in the Bangla text corpus (Dash 2006: 18).

There may, however, arise some problems in this case. It may happen that the most recurrent spelling is not linguistically correct or acceptable. For instance, although the spelling *khet* 'field' is the most frequent spelling used in the Bangla corpus, it is not the correct spelling. The linguistically correct form is *kṣet*. So, here at least, the linguistically correct form should get priority over the frequently used form. Similarly, the form *arun* 'sun' is the most frequent form while the linguistically correct form is *aruṇ*. Also, the most frequent form is *tarun* 'young,' while the linguistically valid form is *taruṇ*. In such contexts, we have to decide whether we shall rely on the frequency of occurrence of a spelling or on the linguistically correct form. It

is at our discretion. However, in our argument, in cases where there is no conflict between high frequency and linguistic correctness, we should rely on frequency. Only in those cases, where there is a conflict between the two options, we may opt for the linguistically correct form.

After selection of a particular spelling, if scope and goal of a dictionary permits, we may furnish all alternative spelling variations in the dictionary. This will increase acceptability and reliability of the dictionary among language users.

8.7 Part-of-Speech

A lexical item, in its free state, has no part-of-speech. It is a linguistic unit waiting to be used in the text. When it is used in a text, it takes a part-of-speech. Thus, part-of-speech of a word is determined from its actual occurrence in a language and not just from its presence in language. Traditional dictionary makers often ignore this fact to provide part-of-speech information to words. How this information is collected is never explicated by them. It is assumed that they collect this information either from their predecessors or observation of usage of lexical items in citation slips. Moreover, their native language intuition and linguistic expertise help them determine the part-of-speech of words. The process of information collection and use is not reliable as it does not give assertion whether—

(a) the information of part-of-speech is validated by an actual example of word use,
(b) the part-of-speech tagged to a word is actually noted in use in a text, and
(c) the number of part-of-speech identified for a word is fixed.

The contribution of a corpus in determining part-of-speech of words is immense. When there is a conflict in the identification of actual part-of-speech of a word, we can directly refer to a corpus, as it provides the actual information of contexts from which we can reliably determine the actual part-of-speech of a word.

Part-of-speech of a word can change while it is put within a particular context. This may happen not only to non-inflected and derived words but also for inflected words. Therefore, it is necessary for us to refer to actual usages of words in contexts and thus determine their parts-of-speech. This function interpretation method is far more useful for dictionary users and language learners since they get a scope to know the part-of-speech of words from the perspective of their use in texts.

Reference to a corpus also helps us to arrange headwords in accordance with their frequency of use in different parts-of-speech. This is never practiced by traditional dictionary makers. Usually, whenever they identify words that are potential to be used in multiple parts-of-speech, they put them in a random sequential order without specifying their frequency of use in different part-of-speech. For instance, in Bangla dictionary, the word *bhāla* ‘good’ is mentioned to be used as an adjective, as a noun, and as an indeclinable. But we do not know in which part-of-speech it registers the highest occurrence.

Using information from corpus, we can solve this problem. We can calculate the frequency of use of a word in particular part-of-speech and furnish the most frequent one in the first position in a dictionary followed by others in sequential order. For instance, if *bhāla* as an adjective is most frequent in use, it will come first followed by its use as a noun and as an indeclinable, in the order of its frequency of use. This will help dictionary users to know in which part-of-speech a word is most recurrent in a language as well as in which part-of-speech it has the least frequency.

8.8 Grammatical Information

Grammatical information of headwords is another important area where a language corpus has direct functional relevance. Grammatical information helps us identify the surface form and meaning of the lexical units. The process of providing grammatical information to words in a dictionary is based on three properties:

(a) The amount of grammatical information,
(b) The type of grammatical information, and
(c) The method of presentation of grammatical information.

In all three areas, we can retrieve all relevant information from a corpus. Since the basic purpose of providing grammatical information in a dictionary is to indicate the lexicosyntactic features of words, we cannot ignore the value of functional peculiarities of words noted in a corpus. When we refer to a corpus for grammatical information, we find morphological and syntactic information from usages of words in texts. This is related either to irregular and unpredictable forms of lexical units or have a direct relation to their syntactical functions to highlight their unique functional roles in rare situations.

Another purpose of grammatical information is to provide clues to contextual meanings of words. It is often noted that actual etymological meaning of words is often changed depending on their use in different contexts. This becomes vital in furnishing semantic information of words in a dictionary, and we can, with the help of concordance method, retrieve the entire range of sense variation of words to capture their actual as well as contextual meanings for a dictionary. In fact, grammatical information becomes the first clue for understanding the meaning and function of words in a language.

When we build a dictionary with reference to a corpus, we observe all the grammatical functions of the words in multiple contextual settings and identify their multiple grammatical categories along with their frequency of use in different grammatical categories. Finally, after summarizing the varieties into several patterns, we furnish information in a dictionary to highlight various types of grammatical information of the headwords.

In case of a digital dictionary, we can follow a different method for furnishing grammatical information of words in the dictionary that may differ from a printed one. In case of a printed dictionary, the normal method of providing grammatical

information is to put this in the form of **labels** (as abbreviated forms of different parts-of-speech) with the headwords. In case of a digital dictionary, we can include both grammatical and semantic information clubbed together and furnish it within a separate frame or tier for each headword.

8.9 Definition and Description

Definition and description of headwords are the most complicated parts of a dictionary since a proper understanding of words on the part of the dictionary users heavily depend on a load of information provided in definition and description. Since these are basic properties of a dictionary, we have to be accurate in the description so that it becomes useful for understanding the meanings of the headwords. We, therefore, argue that a general reference dictionary should have definitional meanings rather than meaning and definition given separately. That means the definition of the headwords should combine both description and meaning.

Definition of words is usually related to their intralinguistic (i.e., lexical) and extralinguistic (i.e., encyclopedic) entities. And this two-level information is a possibility to extract from a corpus by linking their usage in several text types. Also, various semantic aspects of the words may be put together to define the headwords to the maximum satisfaction of end users. This is supported by scholars with the argument that it is necessary to interpret the verbal signs by means of other verbal signs of the same language for their better comprehension (Jacobson 1959: 253).

It is always better to extract information from a corpus to define words in multiple ways as defining a word from multiple angles gives the better treatment a word deserves and a better understanding of the end users. That means multiway interpretation provides better scope to the dictionary users to understand multiple linguistic roles of the headwords in the language. In support of this scholars have argued, 'There should be no one way to define a word; in defining (within certain limitations), the end justifies the means and the end should always be to convey, as accurately and succinctly as possible, the sense of the word being defined' (Urdang 1913: 587).

8.10 Semantic Information

In general, the meaning of a lexical unit is the sum total of its senses used in a language. We need to define the meaning of a headword in such a way that we are able to encompass all possible senses the headword denotes. Since there is no scope for personal whims, we have to define a word in the way we find its use in a corpus. Moreover, since our basic goal is not to prescribe but to describe words, we are bound to describe words as they exist in a corpus text and not as they should be. In this case, we strongly support Hayakawa (1941: 55) who argues, 'The writing of a dictionary is therefore not a task of setting up authoritative statements about the true meanings

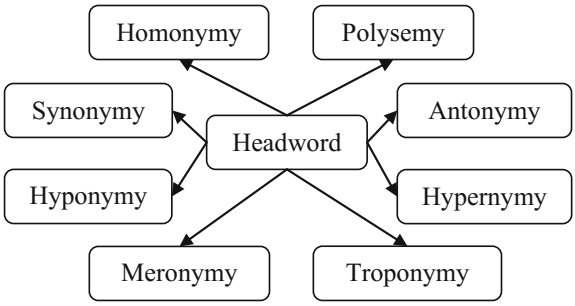

Fig. 8.3 Semantic relation of a headword with lexical semantics

of words, but a task of recording, to the best of ones' ability what various words have meant to authors in the distant or immediate past. The writer of a dictionary is a historian, not a law giver' (Hayakawa 1941: 55).

By analyzing results obtained from a corpus, we can determine the relative frequencies of meanings, which we can furnish to form definitions of the headwords. Since it is not always possible to distinguish frequencies of the meaning of words by analyzing wide varieties of senses expressed by the words, we can perhaps use our linguistic expertise to distinguish among the senses to put them in a certain order in the dictionary. If we find that a word is used in a particular sense exclusively in certain texts types, we may mention this in the dictionary and highlight this domain-specific meaning with special reference to the text type in the corpus.

It is quite normal to find some words exclusively used in texts relating to travel, business, technology, medicine, advertisement, legal documents, etc. It is an essential attribute of a dictionary to specify domains of use of particular words with reference to the citations available in corpora. Similarly, it is possible to link up the use of headwords to children and women or correlate to age, sex, profession, social class, and other demographic factors of language users (Rundell 1996).

Finally, we should also provide detailed information about the network of semantic relation a headword possesses with other words in the language. That means semantic information of a headword should include information of its synonymy, antonymy, polysemy, hypernymy, hyponymy, meronymy, and troponymy, wherever possible, in the dictionary. In essence, a major portion of lexical semantics should come into the dictionary while we try to define semantic information of a headword, as the following diagram shows (Fig. 8.3).

8.11 Usage

Information about the usage of headwords is another important feature of a general reference dictionary without which it becomes useless to many users. As users of a dictionary, we would like to know not only pronunciation, spelling, or meaning of a headword but also its patterns of use in texts. Unless we have clear ideas about the

usage of words, our knowledge about the words is incomplete and partial. Therefore, it is our obligation to provide authentic information about the patterns of uses of words—if possible in all their contextual variations.

We can take help from earlier dictionaries to supply examples of usage of words. Earlier dictionaries, however, cite examples from old and earlier texts. On the other hand, knowledge of dictionary makers is not much reliable for supplying multiple examples of usage variation of words without proper contextual references. To know the usage variation of headwords, therefore, a corpus is the only reliable and authentic source, which we may use effectively to achieve our goal.

A multitextual and multidimensional monitor corpus is the most authentic and reliable source of information about usage variations of words. We should always rely on such a monitor corpus for this purpose. A monitor corpus has strong functional importance in the supply of both diachronic and synchronic information of usage of words with regard to sociocultural dimension. Since this corpus is diachronic in content and composition, information derived from this corpus is highly useful to trace the changes in usage of words as well as to trace the changes in meaning and senses of the words due to usage variation across spatiotemporal factors. In general, the feature of usage of words depends on the following three questions: user, topic, and sense.

(a) Who uses it? (The User)
(b) Where does he use it? (The Topic)
(c) In which sense does he use it? (The sense)

Until and unless satisfactory answers are received for the above three questions, the referential value of a general reference dictionary is greatly reduced and the credit of a dictionary maker is largely diminished. A monitor corpus can address all these questions quite satisfactorily. With reference to the extratextual annotation (stored in a Header File) tagged with texts, we can find out who has used the word. On the other hand, with reference to the text in the corpus we can identify in which domains the word is used; finally, with the help of a concordance list, we can find in which senses the word is used and in how many senses it is used in the language. For instance, consider the concordance list of *that* (Table 8.2) which shows how the word denotes different senses and in which sense it is most frequently used in English.

Finally, it is possible to link up special usage of words to people of various sociocultural and/or professional backgrounds to correlate to age, sex, occupation, ethnicity, social class, education, and other demographic variables. For instance, it is noted that in *British English*, female members use *gorgeous* three times more than their counterparts, and *ever so nice* is typically linked to the 'woman over forty-five, and hardly ever by men of any age' (Rundell 1996). Similarly, we can relate how three quite frequently used idiomatic expressions like *hajabarala* 'haphazardly,' *śabdakalpadrum* 'word as a tree,' and *āboltābol* 'nonsense' are linked to Sukumar Ray—one of the greatest literary figures of Bengal.

Table 8.2 Concordance of *that* to trace usage and sense variations

1	among the myriads of men	that	existed who would pit
2	his angel mother! Who	that	had seen him bright and
3	like a celestial spirit	that	has a halo around him, with
4	and the dearer friends	that	inhabit it. I am surround
5	I inhabit and by you	that	made me, that with the co
6	so quit the sight of those	that	remained to me, and above
7	pain and horror. The poor	that	stopped at their door were
8	gain from any other being	that	wore the human form. “My
9	has found every instance of	that	word in the text
10	Concordance is a program	that	scans a text file and outputs
11	One can perhaps note	that	the line numbers are counted
12	percentage of tokens found	that	were relevant
13	This is the sort of study	that	could have been carried out
14	may be a significant variable	that	is choice of adjective will be
15	Therefore, we believe	that	whatever is said about words

8.12 The Realization

It is clear that in this digital age, a dictionary maker can compile a much better dictionary with the fruitful utilization of data and information obtained from a corpus. In fact, a multitext corpus is largely useful for better representation of linguistic and non-linguistic information of the lexical items selected for inclusion in a dictionary. This trend is becoming effective not only for general reference dictionary but also for other types of dictionary including special and learner dictionary. Availability of corpus has opened up scopes for developing new types of the dictionary which can serve multiple purposes of the users. For instance, the recent trend for designing digital dictionary as a stand-alone resource for various academic and non-academic works is considered as an offshoot of language technology.

Attempts are also made to develop digital dictionaries of scientific and technical terms, synonyms, antonyms, homonyms, polysemy, foreign words, native words, idioms, phrasal units, proverbs, spelling, usage, proper names, person and item names, obsolete words, neologism, slang, cants, etc. In most cases, both text and speech corpora are used for such dictionaries. A speech corpus is used especially for designing dictionaries of spoken texts and of pronunciation. Since these dictionaries are designed with data and information collected from a corpus of present-day language, these are far more informative and reliable. The diagram presented below gives an idea how a language corpus is relevant in proving data and information for compiling a useful dictionary in a language (Fig. 8.4).

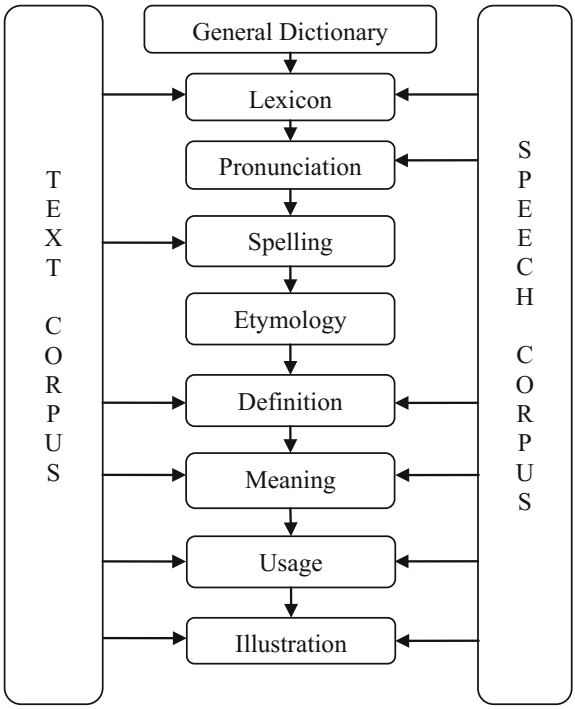

Fig. 8.4 Contribution of corpora in the compilation of a dictionary

8.13 Conclusion

There is scarcely any area of dictionary making where a corpus cannot provide important evidence or data to the lexicographer. In reality, a corpus is of great value in deciding on the word list and on the form of each entry word to be selected for a dictionary. A corpus is also important in defining, first in determining the sense breakdown if the word is polysemous and then in discovering more particularly how best to define each sense with reference to examples culled from a corpus.

It is always good to define words with authentic examples of how such words are used in texts. In this case, at least, a corpus gives extremely good coverage of common words, including common phrasal verbs and idioms. And if a corpus is varied and large, it can also serve to catch and cover many uncommon words. In the act of dictionary making, in essence, a corpus is a resource that has breathed new life into the art of lexicography. 'Wrong decisions are still being made, of course, and always will be, but for the first time lexicographers at least have a sound basis for making decisions and can no longer plead ignorance' (Landau 2001: 305).

The present scenario of success in corpus-based dictionary making leads us to believe that we can also develop something very much resourceful like this for the

Indian languages. The emergence of the corpus has the potential to revolutionize the entire profession of dictionary making as well as the entire world linguistic knowledge storage, management and representation in the panorama of language learning, language education and popular language-related services.

References

Atkins, S. 1998. Starting Where the Dictionaries Stop: The Challenge of Corpus Lexicography. In *Computational Approaches to the Lexicon*, ed. S. Atkins and A. Zampoli, 349–393. Oxford: Oxford University Press.

Atkins, S., and M. Levin. 1995. Building on a Corpus: A Linguistic and Lexicographical Look at Some Near-Synonyms. *International Journal of Lexicography* 8 (2): 85–114.

Atkins, S., and A. Zampoli (eds.). 1998. *Computational Approaches to the Lexicon*. Oxford: Oxford University Press.

Boguraev, B., and J. Pustejvsky (eds.). 1996. *Corpus Processing for Lexical Acquisition*. Cambridge, MA: MIT Press.

Dash, N.S. 2006. *Bahurupi Bangla Banan (Multifaceted Bangla Spelling)*. Kolkata: Daksha Bharati.

Grant, L.E. 2003. *A Corpus-Based Investigation of Idiomatic Multiword Units*. Unpublished Doctoral Dissertation submitted to the Victoria University of Wellington, New Zealand.

Hayakawa, S.I. 1941. *Language in Thought and Action*. New York: Harcourt Brace and World Inc.

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

Jakobson, R. 1959. On Linguistic Aspects of Translation. In *On Translation*, ed. R.A. Brower, 232–239. Cambridge, Mass: Harvard University Press.

Landau, S.I. 2001. *Dictionaries: The Art and Craft of Lexicography*, 2nd ed. Cambridge: Cambridge University Press.

McEnery, T., and R. Xiao. 2010. What Corpora can Offer in Language Teaching and Learning? In *Handbook of Research in Second Language Teaching and Learning*, vol. 2, ed. E. Hinkel, 364–380. London, New York: Routledge.

Moon, R. 1998. *Fixed Expressions and Idioms in English: A Corpus-Based Approach*. Oxford: Clarendon Press.

Rundell, M. 1996. The Corpus of the Future and the Future of the Corpus. Invited talk delivered at a special *Conference on the New Trends in Reference Science* at Exeter, UK (a hand out).

Rundell, M., and P. Stock. 1992. The Corpus Revolution. *English Today* 25: 12.

Samarin, W.J. 1967. *Field Linguistics*. New York: Holt, Reinhart, and Winston.

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

Urdang, L. 1913. Review of Problems in Lexicography. *Language* 39: 586–594.

Web Links

http://www.pearsonlongman.com/dictionaries/pdfs/corpus-lexicography.pdf.

http://grammar.about.com/od/c/g/Corpus-Lexicography.htm.

http://www.slideshare.net/IhsanIbadurrahman/lexicography-engl-6203-third-assignment-draftedited2.

http://www.macmillandictionaries.com/features/from-corpus-to-dictionary/.

http://www.kuleuven.be/grelep/publicat/verlinde.pdf.

http://www.jostrans.org/issue17/art_rogers.pdf.

http://www.euralex.org/elx_proceedings/Euralex1996_1/029_Rosamund%20Moon%20-Data,%20Description,%20and%20Idioms%20in%20Corpus%20Lexicography.pdf.
http://onlinelibrary.wiley.com/doi/10.1111/j.1473-4192.1999.tb00178.x/pdf.
http://donelaitis.vdu.lt/publikacijos/vilniausk.htm.

Chapter 9
Corpus and Dialect Study

Abstract In the present Indian context, we find that many minority language communities are living in different sociocultural and geoclimatic regions across the country. Any kind of systematic study on these languages requires well-formed and properly representative dialect corpus in digital form because a dialect corpus due to its overall representation of the dialect in question is the most reliable resource for studying the dialects in a faithful manner. It is the dialect corpus, and not a general corpus of a standard language, and is of primary importance here, as a dialect corpus is a unique kind of resource that can supply the most authentic data and information of dialect variation that can be investigated with empirical details and verifiable authenticity. Keeping this aspect in view, in this chapter, we shall try to discuss how the study of dialects can be more rational, reliable, authentic, and useful if we carry out research on dialects with the direct utilization of dialect corpus developed in digital form. In this case, the modern dialectologists are not bound to depend, like traditional dialectologists, on data, examples, and information elicited in a controlled manner by analyzing responses elicited from a set of selected informants against the questions asked to them. In this new method, we argue that if we can develop a full-fledged large, multidimensional, and widely representative dialect corpus following the methods, methods and strategies used in modern corpus linguistics, a dialect corpus will be more rational in demographic sampling, more reliable in text representation, more authentic in linguistic observation, and more verifiable in inference deduction.

Keywords Dialectology · Transition · Dialect corpus · Contribution · Relevance Limitations · Field linguistics · Language documentation · Language digitization Language planning · Community development

9.1 Introduction

The compound 'dialect study' refers to the activities of collecting language data from a dialect, analyzing its linguistic features and properties, and developing linguistic

N. S. Dash and L. Ramamoorthy, *Utility and Application of Language Corpora*,
https://doi.org/10.1007/978-981-13-1801-6_9

resources for the growth and development of the dialect. In the present global context of linguistic imperialism and marginalization of minority languages, the activities relating to digitization and documentation of minority languages have become issues of great international interest, since proper restoration of nearly extinct and endangered languages is linked with preservation of knowledge, history, and heritage as well as preservation and development of culture of the speech communities. In a country like India, where thousands of minority languages are blinking at the verge of extinction, it is absolutely necessary to develop methods and strategies for documentation and preservation history, knowledge, and heritage of these languages in digital form (Dash 2005). Through this process, we shall not only collect language data but also collect information relating to life, culture, history, heritage, customs, ethnicity, mythology, etc., of the minority language communities and preserve them as valuable resources to be used in language planning and language rejuvenation. Thus, dialect study will become an integrated part of our social responsibility where each and every Indian linguist should commit some time of his or her academic life for the growth and advancement of minority Indian languages.

In recent years, dialect study has shifted its focus from dialect data to dialect corpus. A dialect corpus generated in digital form has become an indispensable resource for clinical analysis of a dialect (Austin and Sallabank 2011). This new trend of dialect research promises to flourish in coming years as modern dialectologists are gradually shifting their attention from manually compiled limited dialect databases to a systematically compiled large dialect corpus for authentic data, reliable information, and appropriate examples (Lindström and Pajusalu 2002; Grenoble and Furbee 2010). With the advent of new technology and strategies for collection, processing, analysis, and utilization of a dialect corpus, the field of dialect study acquires a new dimension in dialectology as a dialect corpus grows in form and content to provide more dependable evidence for unique elements, features, and information of dialects with full empirical details and verifiable authenticity (Peitsara 2000).

Within the traditional frame of dialect study, a person (known as a dialectologist) is usually engaged in the collection, storage, and analysis of dialect data. He is normally interested to study how dialects vary across regions; how dialects differ from each other despite geographical proximity; how dialects are spoken in various geographical regions; how dialects vary with regard to a standard variety; how a dialect moves toward the state of standardization, convergence, or extinction; how a dialect borrows from other dialects; and similar other questions. For centuries, these have been some of the primary queries in dialect study, and due to this reason dialect study has evolved as one of the empirical fields of applied linguistics.

The modern concept of *dialect study* is, however, modified to a great extent to include many new queries within its frame. Besides including the goals of traditional dialect study, it also proposes to do the following things:

(1) Collect dialect data and dialect-related information from various dialects and regional varieties.
(2) Store these data systematically in the digital archive with extra- and intra-textual annotation.

(3) Process these data to make these usable by man and machine.
(4) Analyze these data and draw inferences from the analysis.
(5) Develop linguistic resources for dialect communities.
(6) Utilize these resources for restoration and preservation of dialects.
(7) Provide data and information for language planning and language revival.

All the proposed tasks may be carried out with an exhaustive analysis of large and varied dialect corpus developed with real-life evidence of dialect use. In essence, dialect study should necessarily be based on corpus for its effective result (Austin and Sallabank 2014). And for access to a dialect corpus, we require special equipment, hardware, adequate storage capacity, processors to handle large text and audio files, software, a strong operating system for handling data, display and editing facilities, transcription modalities of audio recordings, acoustic analysis of speech data, etc.

In this chapter, we propose for developing a dialect corpus made with the adequate amount of dialect data and dialect-related information directly collected from the speakers in their own sociocultural and geoclimatic settings through face-to-face interviews. These interviews will not only have responses based on previously prepared questionnaires but also include several thoughtfully designed free discourse text (FDT) for capturing all possible shades and shadows of life and living of the target dialect community. A dialect corpus developed in this manner will be authentic, representative, informative, and reliable as the data is collected in the mode of free discourse texts with close reference to the existence of speakers in their own home environment (Hinton et al. 2017).

In Sect. 9.2, we discuss the nature of transition in approaches to dialect study from traditional practice to modern digital method; in Sect. 9.3, we redefine dialect study in light of modern dialect corpus; in Sect. 9.4, we present a theoretical frame and content structure of a dialect corpus; in Sect. 9.5, we discuss the contribution of a dialect corpus in dialect study; in Sect. 9.6, we highlight the relevance of a dialect corpus in linguistics and other disciplines; in Sect. 9.7, we address some of the limitations of a dialect corpus.

9.2 The Transition

For several decades, we have noted that dialectologists are interested to know how dialects vary across geographical regions. To achieve this goal, they have tried to analyze limited set of manually collected dialect data to know how different dialects are spoken at different geographical regions, how dialects deviate from standard varieties, and what kinds of mutual linguistic interrelation can be recorded between the dialects in one hand and with the standard variety on the other. These are, no doubt, important queries, due to which dialect study is considered as one of the primary empirical fields of linguistics. In essence, within the general frame of traditional dialect study, scholars aim at collecting data from dialect varieties, analyze them, and draw some inferences based on analyzed data. In most cases, these studies

are carried out on controlled data samples rather than using large dialect corpus compiled in a scientific manner. This traditional method posits little problem to the traditional dialect investigators since their primary goal is to focus on the vocabulary and pronunciation of words used by particular dialects rather than focusing on other properties of dialects relating to morphology, syntax, semantics, and ethnology.

As modern dialect investigators, we have different goals from that of traditional dialectologists. We are not much interested to know how dialects vary across geographical regions. Rather, we are more interested to identify how information and knowledge are encoded within dialects and how these can be faithfully extracted and utilized for global access. The availability of multidimensional digital dialect corpus in recent times has opened up many new methodical as well as directional changes in dialect study. Several dialect corpora which are now available for global access (e.g., *Nordic Dialect Corpus*, *Freiburg English Dialect Corpus*, *Helsinki Corpus of British English Dialects*, *Estonian Dialect Corpus*) contain systematically designed dialect samples which adequately represent the dialects from which these are developed. Moreover, all major language archives and data centers such as *the Oxford Text Archive, the Bank of English, the British National Corpus, the American National Corpus, the Linguistic Data Consortium, the Spanish Speech Corpus, the Penn-Helsinki Language Database* have collected large and widely representative corpus of dialects spoken at various regions of Europe and America. These dialect archives have generated the momentum to inspire many others to produce dialect corpus of different types (e.g., the *Syntactic Atlas of the Dutch Dialects*) with the direct utilization of tools and techniques used in modern corpus linguistics (Barbiers et al. 2005, 2008). In essence, the new generation of scholars of dialect study has adopted a new approach for generating a new kind of dialect corpus, which has opened new scopes for multidimensional analysis of dialects which have never possible before.

In principle, a particular dialect, due to its geographical, typological, and genealogical proximity, may be linked up with the standard variety as well as with other regional varieties. This makes a dialect an important area of investigation within the broad spectrum of sociolinguistics which tries to understand the hidden network of the interface between language, speakers, community, and culture on the axis of time and space (Fig. 9.1). Since the study of a dialect is actually the study of a language variety used by a particular speech community at a particular geographical location at a particular point in time, it cannot be complete without reference to the culture, the society, and the community who uses the dialect. That means, the multifaceted spectrum of life and living of community has to be accurately manifested in the corpus that wants to represent the dialect.

This leads us to argue that the analysis of a dialect and its speakers cannot be complete without the analysis of widely representative and multidimensionally developed dialect corpus made with a large number of representative samples of texts the speakers of the community actually use in their daily course of living. Therefore, we argue for generating multidimensional dialect corpus in digital form, which can be used to extract necessary data, examples, and information to understand various aspects of the life of the speakers. Due to the compositional width and representational diversity, a dialect corpus can supply necessary information about various linguistic

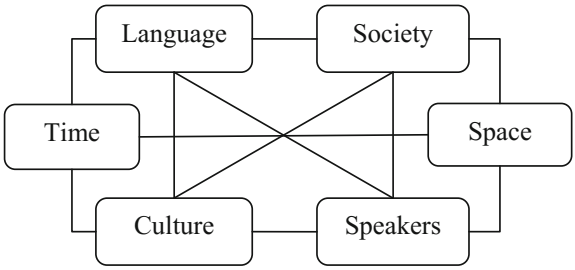

Fig. 9.1 Dialect within the broad spectrum of sociolinguistics

and extralinguistic aspects of the dialect to formulate qualitative-cum-quantitative observations about the community (Gippert et al. 2006).

The traditional methods of dialect study, although useful to a certain level (Samarin 1967), have many limitations, which are not possible to overcome if we do not adopt the new methods of dialect data collection and analysis. Within traditional scheme, notable limitations are observed in the preparation of questionnaire, selection of informants, the manner of conducting interviews, elicitation of responses from informants, recording and collecting responses, processing databases, interpretation and analysis of data, and in drawing resultant inferences. That means, at every stage of traditional dialect study, the dialect researchers have been crippled with several practical limitations, which forced them to draw skewed observations and questionable inferences. Perhaps, the non-availability of advanced tools and techniques in dialect data collection, processing, and analysis has been one of the strong barriers in traditional dialect study—as far as the studies of the earlier dialectologists are concerned.

In this context, we argue that the advanced tools and techniques that we use in corpus linguistics can be highly beneficial for modern dialect study. We can use these techniques to overcome the problems relating to skewedness in data elicitation, an imbalance in text representation, errors in text processing, mistakes in text analysis, and faults in inference deduction. Moreover, by applying the advanced statistical and computational tools on digitized dialect corpus we can infer more reliable and authentic observations that are open to all kinds of empirical verification. We can have a better understanding of the nature of the transition from traditional dialect study to modern dialect study with reference to the following list (Table 9.1).

For modern dialect researchers, it is comparatively an easy task to collect dialect data in digital form in a comprehensive manner with samples of texts taken from all kinds of linguistic interaction directly from a speech community, while the native speakers are involved in various kinds of regular dialogic interactions. The impromptu representation of texts of regular verbal interactions will make a dialect corpus largely balanced and reliable to overcome the problems of skewedness and imbalance with regard to a text representation. Moreover, systematic application of statistical and computational methods of text processing and analysis will ensure error-free outputs for the investigators to draw faithful conclusions. Thus, methods and techniques used in corpus linguistics can add a new dimension in the area of dialect study never visualized before.

Table 9.1 Traditional dialect study versus modern dialect study

Traditional dialect study	Modern dialect study
Limited amount of data	Unlimited amount of data
Mostly one-dimensional	Mostly multidimensional
Specific study	Holistic study
Sample-based	Corpus-based
No use of technology	Heavy use of technology
No use of text processing tools	Heavy use of text processing tools
Study how dialects vary across geographical regions	Study how information and knowledge are encoded within dialects
Questionnaire-based elicitation	Mostly free discourse texts
Purpose is documentation	Purpose is restoration, knowledge generation
Non-participation of dialect community	Direct participation of dialect community

9.3 Redefining Dialect Study

In this section, we propose a new method, which we call corpus-based dialect study. In this method, we try to provide a better perspective toward dealing with the problems that are faced by traditional dialect researchers. The basic component of this method is a dialect corpus which is developed in digital form following the methods and principles of corpus linguistics. A dialect corpus developed in digital form contains samples of the text of written and spoken interactions in a faithful and uniform manner. While the written part stores several samples of written text, the spoken part stores samples of spoken text collected from various spoken interactions. Proportional representation of both types of text becomes indispensable in subsequent stages of dialect analysis and interpretation to reflect on the linguistic features of the dialect as well as on the people who use it.

The composition of some dialect corpora recently developed in digital form provides us necessary insights into how we can capture the representative text samples from a dialect (Francis 1980). Also, it provides us necessary guidelines for analyzing a dialect corpus by which we can faithfully reflect on the special features of a dialect. A digital dialect corpus due to its variety in texts, multidimensionality in content, and diversity in form excels over the traditional dialect databases compiled manually, as a digital dialect corpus captures the overall usage variation of a dialect with a focus on its diversity and variety. Therefore, to know the life and society of the people through the dialect they use, to extract the knowledge and information embedded within a dialect, as well as to identify the finer traits of difference underlying between a dialect and a standard variety, it is rational to depend on digitally developed multidimensional dialect corpus than on one-dimensional dialect data collected manually for object-oriented studies.

A digitally developed dialect corpus is more faithful in supplying necessary linguistic information and data relating to phonetics, phonology, morphology, lexicon,

semantics, syntax, and other linguistic properties of a dialect. For instance, surreptitiously collected normal speech corpus is highly useful in providing necessary segmental, suprasegmental, and phonological information to study the phonemes, allophones, sound systems, speech patterns, intonations, and other aspects of speech used in a dialect. We can infer important statistical results from a speech corpus on the varieties of use of various sound elements based on which we can make individual as well as general remarks about the usage patterns of sound segments in a dialect. And such a study on a dialect is necessary to know how a dialect differs from its sister dialects as well as from the standard variety. For instance, data and information retrieved from the *Helsinki Corpus of British English Dialects* are used by investigators to mark the phonological similarities and differences among the English dialects as well as between the dialects in one hand and the standard variety on the other (McEnery and Wilson 1996: 110).

As modern dialect researchers, we are interested in using techniques of corpus linguistics within the main frame of our activities as we are willing to go beyond the sphere of general observation of dialect elements to the realm of 'scientific truth' which corpus linguistics always tries to uphold. We desire while we analyze a dialect corpus, to overcome the pre-defined boundaries of general concerns of traditional dialect study to delve into the inner sides of words and meanings, pronunciations, and other regular aspects of dialects. Moreover, we want to delve into various sociolinguistic issues like gender, occupation, ethnography, discourse, pragmatics, aesthetics, ecology, networking, heritage, history, culture, and hosts of other aspects to look at a dialect not as an isolated phenomenon but as a vibrant social entity directly linked with several extralinguistic controlling factors. We approach these issues as important areas of dialect study and treat these in pure empirical light to understand a dialect and its speakers.

Through an empirical analysis of dialect corpus, we can show the differences in views of the linguists and non-linguists about a dialect. We can also show the kinds of knowledge the non-linguists usually use to 'know' a dialect. That means, information and data stored in a dialect corpus contain some relevant cues that help us to map the conceptual and cognitive interfaces between a dialect and its users which is termed as 'Perceptual Dialectology' (PD) (Petyt 1983).

The basic argument is that a systematically developed digital dialect corpus is a good source of data for detailed and authentic information about various spheres of life of a speech community. It is perhaps the most attractive area of interest for both linguists and non-linguists who are willing to track the divergences and convergences of various routes of life and living of the dialect speakers (Peitsara 1996). As a result, linguists, sociologists, anthropologists, and social psychologists who deal with the interfaces of language and life can easily find the wealth of authentic data to write a more general account about dialects with a focus on the ethnography of the dialect community. As modern dialect researchers, we are interested in a holistic model and willing to relate language with oral history, performance studies, psychol-

	Written Text Samples	Spoken Text Samples
Imaginative texts	mythology, legends, history, geography, folktales, folklore, fairy tales, fables, general stories, ghost stories, love stories, rhymes, riddles, songs, ballads, proverbs, idioms, poems, plays, elegies, etc.	folksongs, folktales, lullabies, fairy tales, oral stories, tales, riddles, fables, rhymes, ballads, elegies, poems, ghost stories, love stories, songs, proverbs, idioms, poems, plays, elegies, puzzles, etc.
Informative texts	Business and commerce, social life, history, religion, faiths, cults, rituals, nature, politics, culture, environment, literature, practice, norms, agriculture, customs, feasts, festivals, games, sports, traditions, health, professions, hygiene, cultivation, migration, etc.	business, agriculture, religion, migration, environment, history, nature, faiths, geography, politics, norms, social rules, systems, cults, traditions, customs, socialization, rituals, culture, festivals, folk science, health, games, sports, hygiene, ailments, etc.

Fig. 9.2 A general composite structure of a dialect corpus

ogy, religion, belief, semiotics, sociology, culture, gender, aesthetics, dance, music, narrative and verbal arts, linguistics, ethnomusicology, the literature, anthropology, history, heritage, ecology, area study, folk art, communication politics, rituals, festivals, humor, identity, science, food, health, medicine, art, and all other issues of a dialect community.

9.4 Structure of a Dialect Corpus

In principle, the structure of a dialect corpus, which we propose to design following the methods of corpus linguistics, should contain representative texts from both written and spoken sources. While the written text samples should be compiled from written texts available in a dialect, the spoken text samples should be collected from normal dialogic interactions that take place within a sociocultural spectrum of linguistic interactions of a dialect community. In the following diagram (Fig. 9.2), we try to show a general frame for developing a dialect corpus in accordance with the scheme of a digital corpus generation.

According to this scheme, a representative dialect corpus is made up of two parts: (a) written text samples and (b) spoken text samples. Each part is constituted with imaginative text samples and informative text samples as the above diagram shows. With the text samples gathered from the variety of sources, a dialect corpus represents a ‘composite structure’ of a dialect taking into account the width and variety of life and society of the people belonging to the dialect. The fields listed in the list above show that while the written part preserves samples of written texts collected from varieties of sources (wonder, if all these are available in a dialect), the speech part contains samples of spoken texts from daily dialogic interactions relating to various aspects and issues of life. The written text samples, if necessary, may be

further augmented with new data collected from archives or published sources or transcribing the spoken texts produced by the informants into written form.

On the other hand, the spoken text samples can be collected from innumerable contexts of dialogic interactions with the speakers of the dialect covering all possible aspects, issues, and events that take place in the life of the community members. The spoken texts may be rendered into written form by using the processes of phonetic transcription and orthographic annotation. Thus by way of systematic collection, we can develop a good, balanced, and properly representative dialect corpus. We admit that the task of compiling such a dialect corpus is quite difficult and time-consuming, as it asks for long-term planning, large-scale investment, collective enterprise, and fruitful utilization of the methods of corpus linguistics. However, once we succeed in developing a dialect corpus of this kind, we add up a new dimension to the field of dialect study to make it far more effective and reliable. Moreover, by systematic analysis of data stored in a dialect corpus, we can produce new evidence to look at a dialect from a new perspective.

9.5 Contribution of a Dialect Corpus

In the diagram below (Fig. 9.3), we propose a scheme to show how linguistic as well as extralinguistic data and information may be extracted from a dialect corpus to be used in various kinds of research and development work. When we delve into the diagram, we find that a raw dialect corpus may be made up with samples of spoken texts of various kinds of a dialect. And these texts may be collected from a dialect community in a systematic manner following the rules and methods of digital corpus generation to make the dialect corpus maximally varied and widely representative of the target dialect community.

After a raw dialect corpus is collected in electronic form, it is passed through several stages of corpus processing that involves indexing of text samples, text-type categorization, transcription of spoken texts, type–token analysis, lexical division, frequency count, prosodic annotation, semantic annotation, grammatical annotation, concordance, lemmatization, local word grouping, key-word-in-context, collocation, parsing. Most of these works are carried out with the help of corpus processing tools and systems developed for processing a general language corpus. A processed dialect corpus is further classified to produce a core dialect corpus which contains only the indispensable linguistic and extralinguistic data and information relating to a dialect. Now, we have, at their disposal, a core dialect corpus which is designed systematically to serve our all linguistic purposes. From this corpus, we can retrieve data, examples, and information of various types to address various research and application needs of a dialect.

The most important part is related to the extraction of information and data of various types from the core dialect corpus. People of various domains and interests can utilize this corpus to address their varied needs. For instance, phoneticians can extract sounds, phonemes, allophones, and all kinds of phonological data and information;

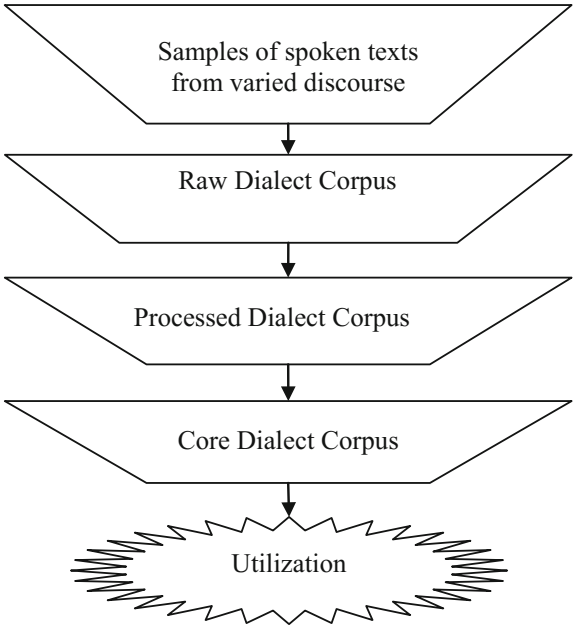

Fig. 9.3 Utilization of a dialect corpus in dialect study

grammarians and morphologists can collect morphs, words, and related elements; semanticists can retrieve necessary information of word meanings; syntacticians can collect elaborated data and information about phrase, clause, and sentence formation; lexicologists can retrieve lexicon, idioms, phrases, proverbs, and other data to investigate various linguistic issues and events noted in a dialect.

On the other hand, people who are not directly linked with core linguistics or dialectology can also benefit from a core dialect corpus. For instance, scholars of sociolinguistics, anthropology, ecology, sociology, ethnography, and language planning can use data and information from the core dialect corpus to address their queries about various linguistic and extralinguistic issues relating to a dialect as well as about the life, living, and society of the dialect speakers. Thus, a core dialect corpus obtained from a raw dialect corpus can satisfy the requirements of all concerned who want authentic data and information relating to a dialect and its users.

9.6 Relevance of a Dialect Corpus

Perhaps, there should not be any question with regard to the relevance of a dialect corpus in research and development of a dialect (Ihalainen 1991). Since the utility of a dialect corpus in dialect study is self-explanatory, the development of a well-formed and well-framed dialect corpus should be our primary concern. We can visualize the referential relevance of a dialect corpus if we consider dialect study as an important

domain of sociolinguistics where dialect data is perceived as one of the most trusted resources to look into the multidimensional fabric of the life and living of a dialect community. In the following paragraphs, we try to show how data and information obtained from a dialect corpus become valuable inputs for sociolinguistics and other branches of social science directly relating to the life and living of a dialect community (Ihalainen 1994).

A dialect corpus provides not only necessary information relating to sounds and phonemes used in a dialect, but also supplies a huge amount of data relating to morphs, words, idioms, sentences, phrases, and other linguistic elements used in a dialect. This leads us to make a reliable statistical estimation on the patterns of use of various linguistic items in a dialect based on which we can make a general and specific observation about a dialect and its speakers.

We have observed that many linguistic data, information, and properties of a dialect 'die' over a period of time. This is a universal phenomenon noted in all dialects. As a result, the folk texts (e.g., *folk tales*, *fairy tales*, *fables*, *legends*, *stories*, *rhymes*, *lullabies*, *riddles*, *puzzles*), which were once very much vibrant in a dialect, die in forever without any scope for their retrieval. This is an irreparable loss for a dialect community that possessed these as well as for a standard variety to which the dialect belongs. Since these folk texts are an undisputed source of information and knowledge, they contribute in different capacities to shape up life and society of the dialect speakers. A dialect corpus preserves these resources for future use.

The complete picture of the life of a dialect community is never captured unless the samples of both imaginative and informative texts are analyzed together to explore the diversities of life. Since other branches of social science equally benefit from these, the relevance of a dialect corpus is further expanded beyond the realm of dialect study. Therefore, to portray a comprehensive picture of life, living, and culture of a dialect community, we require a large and widely representatively dialect corpus that can help us with data and information for carrying out sociocultural, socioeconomic, and sociopolitical investigations.

A dialect corpus is needed to draw a line of distinction between a dialect and a standard variety. A dialect corpus usually stores a large stock of old words, ethnic terms, specialized dictions, codes, jargons, idioms, phrases, epithets, and proverbs, etc., which become useful in supplying valuable corroborative linguistic data and information to mark the unique linguistic identity of a dialect against the pervasive impact of a standard variety.

A dialect corpus, due to uniqueness in composition and content, is able to contribute valuable linguistic and extralinguistic information and data which are essential in descriptive linguistics, historical linguistics, comparative linguistics, sociolinguistics, and ethnolinguistics.

Traditional dialect studies show, with a handful of data, how the standard pronunciation of words varies in a dialect. In most cases, examples furnished in such studies vary within a small range of citation based on the collection of suitable words by investigators. We can ask questions here regarding the number of citations as well as about the scarcity of evidence. Although people claim that there are differences between a dialect and a standard variety at utterance level, these differences are not

adequately explicit due to lack of large and widely representative data. A dialect corpus can be quite useful here as it provides a huge amount of speech data marked with distinct utterance variations.

Traditional dialect studies claim that a dialect possesses specific sets of affix, postposition, case marker, particles, etc., which are not found in a standard variety. We cannot accept this claim to be right due to the scanty amount of data based on which such claims are made. We may also raise questions regarding the source of data, methods of extraction, and nature of their representative potentiality. These questions will become irrelevant if we make claims with data and examples directly extracted from a dialect corpus. In essence, the lack of a representative dialect corpus may force traditional dialectologists to restrain themselves from furnishing enough evidence for establishing their arguments.

It is often claimed that the grammar of a dialect is different from that of the standard variety. There is, however, no valid proof to validate this observation. Nobody ever made a comparative study in this area by analyzing a dialect corpus with that of a corpus of a standard variety. Also, there has never been any effort to estimate how differences are statistically significant. Answers to these questions may be found only when corpora of both the varieties are properly analyzed, compared, and statistically measured. That means, without reference to corpus whatever is said about the grammar of a dialect and a standard variety is actually a crippled generalization based on partial intuitive observation.

Within the present frame of dialect research, a dialect corpus is indispensable in the act of addressing various issues of dialectology. A dialect corpus is used to study about how a dialect varies across places and times, how the vocabulary of a dialect increases or decreases with the change of time, how the meanings of words change with time and event, how a dialect varies within the same geographical region, how a particular dialect within a group of similar dialect varieties becomes a standard one, how linguistic features of a particular variety contribute in the process of standardization, etc. All such questions can be rightly addressed with direct reference to a dialect corpus.

Finally, those who work in dialect-based studies of life, society, and culture may find a dialect corpus immensely useful for authentic demographic, sociocultural, and geoclimatic data and information. In fact, a dialect corpus is the most faithful resource for them wherefrom they can collect a list of words, specialized terms, idioms, phrases, proverbs, etc., to generate knowledge texts like dictionaries and word books as well as to show the interfaces underlying between a dialect and its speakers.

In recent years, we have observed a close interface between dialect study and formal syntax in which the individual syntactic forms are examined between the related dialects, although no empirical basis is found for studying minimal diversions in which sentences of dialects vary. This implies that the linguistic field studies and the procedures through which the dialect data is elicited require necessary up-gradation to facilitate comparative analysis of this kind. The availability of a dialect corpus can open up new opportunities to explore this area with authentic reference to empirical examples.

9.7 Limitations of a Dialect Corpus

There are some limitations of a dialect corpus, however large and diversified it may be. And due to these limitations, the use of a dialect corpus in the study of a dialect has not been wide and universal. We try to refer to some of the limitations in the following paragraphs.

It is necessary to transcribe the spoken texts that are captured within a dialect corpus into written form to process, analyze, and interpret these. However, during the time of transcription of spoken texts, we invariably lose some amount of phonetic and prosodic information of normal speech. This happens due to complexities and technical constraints involved in the conversion of spoken texts into written form. Since speech and writing (i.e., *acoustic vocalic* vs. *graphic*, *auditive* vs. *visual*, etc.) are different, the spoken texts of a dialect contain some unique features, which are indeed very difficult to represent in the written form (Eggins 1994: 56). For instance, conversion of texts in which speakers are talking together simultaneously is really difficult to represent in linear order within a written text. Naturally, in such cases, the accuracy of transcription depends heavily on the skill of a transcriber as well as on the purpose of a study. For example, if a spoken text is intended for studying the morphosyntactic features of speech, the normal orthographic transcription may be sufficient, but it is not always easy to decide the spelling of certain lexical constructions used in the spoken texts.

Another limitation is the recognition of speech units, the meaningful entities in the speech of individual speakers, which are recorded and put into a dialect corpus. Sentences, although treated as fundamental structural units in formal grammatical descriptions, cannot be easily discerned in dialogic interactions or group conversations (Weigand and Dascal 2001: 15). That means, we do not have reliable cues to identify sentences in spoken texts of dialogues except in terms of their semantic content. Orthographic transcription conventionally contains sentence-final punctuation marks such as periods, full stops, exclamation marks, and question marks. In actuality, these are manually inserted by a transcriber as cues to reflect on the features such as changes of topics, rising and falling intonations, pauses in a stream of speech, and intention of speakers. But spoken texts show that a speech proceeds in a long sequence of paratactic units without any indication for most of the features. The punctuation marks that are used in spoken text transcription are actually a kind of forceful imposition of the rules of standard written text on the structure of a spoken text.

Even if we agree that the characteristic features of informally spoken texts of a dialect are possible to detect in some way or other, we have no cues for those spoken texts that occur in a typical ongoing flow of speech in which both the clausal and non-clausal elements are interfaced together more freely than they are noted in more formal varieties of speech. Since dialect speeches are produced in 'real-time situation' and 'on the spot,' speakers proceed without a premeditated plan and they often change the structure in the middle of a sentence. As a result, dialect speeches become highly punctuated by pauses, hesitations, repetitions, diversions, false starts,

fillers, non-beginnings, non-ends, and similar such elements. Moreover, to save both time and energy, speakers aim at reducing the length of what they have to say by not expressing the words considered dispensable in understanding the texts. This often results in ellipses and tags from the point of view of the standard form of formal spoken texts. Therefore, while analyzing spoken texts of a dialect corpus, it is necessary to treat sentence structures in a larger context of discourse to find out the patterns typical to individual speakers.

The system of analyzing spoken texts included in a dialect corpus is still, to a large extent, based on the method used to analyze a text of a written corpus. We, therefore, have doubts to what extent the techniques and tools that are used to analyze written text corpus can be effective for analyzing spoken text corpus. We also doubt whether we should use the nomenclature of written text analysis (e.g., *sentences*, *clauses*, and *phrase*) for analysis of spoken texts as this nomenclature hardly fit into the frame of spoken text units. On the other hand, we introspect if we should analyze spoken texts at the unit level—at the level of meaningful entities—in the speech of individual speakers. Perhaps, we require further exploration in this area to find out a workable solution.

9.8 Conclusion

Dialect study is an application-oriented field where a dialect corpus is an indispensable resource for linguistic investigation and analysis. In principle, dialect researchers need a dialect corpus of spoken texts not only to study regional uniqueness observed in a variety but also to develop dialect-based resources such as dialect dictionaries, dialect grammars, and word books.

A dialect corpus is different from a corpus made in standard variety in the sense that a dialect corpus is characteristically archaic in nature and less open to change. It becomes useful to the study of those linguistic features and properties, which are considered rare or obsolete in the standard variety. Thus, a dialect corpus, by virtue of its rare linguistic properties, becomes a valuable linguistic treasure for descriptive linguistics, historical linguistics, comparative linguistics, and applied linguistics.

Based on the above observations, we conclude that in a multidialectal country like India, we need to pay attention toward the generation of dialect corpora for the dialects used in India. This will enable us to preserve all possible dialect varieties found in the country. Moreover, analysis of these dialect corpora will yield many new insights and information by which the dialects and their people will be benefitted both linguistically and culturally.

References

Austin, P.K., and J. Sallabank. 2014. *Endangered Languages: Ideologies and Beliefs in Language Documentation and Revitalization*. London: British Academy.

Austin, P.K., and J. Sallabank (eds.). 2011. *The Cambridge Handbook of Endangered Languages*. Cambridge: Cambridge University Press.
Barbiers, S., H. Bennis, G. De Vogelaer, M. Devos, and M.H. van der Ham. 2005. *Syntactic Atlas of the Dutch Dialects*, vol. I. Amsterdam: Amsterdam University Press.
Barbiers, S., J. van der Auwera, H. Bennis, E. Boef, G.D. Vogelaer, and M.H. van der Ham. 2008. *Syntactic Atlas of the Dutch Dialects*, vol. II. Amsterdam: Amsterdam University Press.
Dash, N.S. 2005. A Brief Historical Survey on the Use of Handmade Language Databases in Linguistics Studies. *Language Forum* 31 (1): 17–39.
Eggins, S. 1994. *An Introduction to Systemic Functional Linguistics*. London: Pinter Publishers.
Francis, N. 1980. *Dialectology*. London: Longman.
Gippert, J., N.P. Himmelmann, and U. Mosel (eds.). 2006. *Essentials of Language Documentation*. Berlin: Mouton de Gruyter.
Grenoble, L.A., and N.L. Furbee (eds.). 2010. *Language Documentation: Practices and Values*. Amsterdam: John Benjamins.
Hinton, L., L. Huss, and G. Roche (eds.). 2017. *Routledge Handbook of Language Revitalization*. London: Routledge.
Ihalainen, O. 1991. A Point of Verb Syntax in South-Western British English: An Analysis of a Dialect Continuum. In *English Computer Corpora. Selected Papers and Research*, ed. S. Johansson and A.-B. Stenström, 201–214. Berlin: Mouton de Gruyter.
Ihalainen, O. 1994. The Dialects of England Since 1776. In *The Cambridge History of the English Language. Vol. V: English Language in Britain and Overseas. Origins and Development*, ed. R. Burchfield, 197–274. Cambridge: Cambridge University Press.
Lindström, L., and K. Pajusalu. 2002. Corpus of Estonian Dialects. In *Proceedings of the 11th International Conference on Methods in Dialectology*, 78–82. University of Joensun, Finland, 5–9 Aug 2002.
McEnery, T., and A. Wilson. 1996. *Corpus Linguistics*. Edinburgh: Edinburgh University Press.
Peitsara, K. 1996. Studies on the Structure of the Suffolk Dialect. In *Speech Past and Present. Studies in English Dialectology in Memory of Ossi Ihalainen*, ed. J. Klemola, M. Kytö, and M. Rissanen, 284–307. Frankfurt Main: Peter Lang.
Peitsara, K. 2000. The Prepositions ON and OF in Partitive and Temporal Constructions in British English Dialects. *Neuphilologische Mitteilungen* 101 (2): 263–326.
Petyt, K. 1983. *The Study of Dialect*. London: Andre Deutsch.
Samarin, W.J. 1967. *Field Linguistics: A Guide to Linguistics Fieldwork*. New York: Holt, Rinehart and Winston.
Weigand, E., and M. Dascal (eds.). 2001. *Negotiation and Power in Dialogic Interaction*. Amsterdam: John Benjamins Publishing Co.

Web Links

http://benszm.net/omnibuslit/Anderwald_Szmrecsanyi_HSK_webversion.pdf.
http://gradworks.umi.com/33/70/3370619.html.
http://www.helsinki.fi/varieng/CoRD/corpora/Dialects/.
http://www.lancaster.ac.uk/staff/hollmann/WBH_AS_corpora_Lancs.pdf.
http://www.let.rug.nl/~heeringa/dialectology/thesis/thesis02.pdf.
http://www.murre.ut.ee/estonian-dialect-corpus/.
http://www.tekstlab.uio.no/nota/scandiasyn/.
http://www2.anglistik.uni-freiburg.de/institut/lskortmann/FRED/.
http://wwwling.arts.kuleuven.be/qlvl/prints/.
https://pure.knaw.nl/portal/en/.

Chapter 10
Corpus and Word Sense Disambiguation

Abstract Every natural language has a large set of words, which, when these are used in a piece of text, may vary in sense denotation. It has been noted that for ages that context, where these words are found to be used, can play an explicit and active role to influence the words to deviate from the original sense to generate new senses. And these new senses are usually contextualized or context based. The newly acquired senses often vary from the original senses of words usually derived from their origin or etymology. This phenomenon of words in a natural language has several long-standing problems relating to understanding and cognition of word meanings, using word meanings in machine learning as well as presenting word meanings in the dictionary. In this chapter, we shall describe the 'corpus-based approach' to deal with the phenomenon of sense understanding of words. We shall try to discuss how the information extracted from a corpus can help us attribute meaning of words to their unique distributional information and contextual environments. While distributional information refers to the frequency distribution of the senses of the words, contextualized environments refer to the setting of occurrence of the words in some particular textual situation. We shall also try to show how we can extract necessary information from the contexts of the words used in a corpus to pick up necessary cues for understanding the actual contextualized senses. To substantiate our arguments, we shall try to draw supporting data, information, and evidence from the Bangla corpus of written texts developed in the TDIL project.

Keywords Word meaning · Word sense disambiguation · Issues in sense variation · Context in sense variation · Context · Interface among the contexts Corpus in sense disambiguation

10.1 Introduction

All natural languages have a large set of words, which, when these are used in a piece of text, may vary in a sense denotation. It is the contextual situation that actually plays an explicit (rarely implicit) and active (rarely passive) role to influence the

N. S. Dash and L. Ramamoorthy, *Utility and Application of Language Corpora*,
https://doi.org/10.1007/978-981-13-1801-6_10

words to generate novel contextualized senses, which often vary from the original senses that are derived from etymological information of the words. This implies that context plays a highly crucial role in partial or total modification of senses as well as in projecting new senses of words. Since it is not always possible to decipher the contextualized senses of words just by observing them in simple contextual frames, several intralinguistic as well as extralinguistic factors (e.g., *knowledge of the external world*, *discourse dimensions*, *pragmatic interfaces*, *conceptual interface underlying an act of communication*, *register variables*, *diverse social relations existing among participants*, *sociocultural background of different language acts*, *co-texts*, etc.) contribute toward obtaining the actual sense of words. Therefore, for the purpose of understanding the actual senses of words, it is always necessary to go for a thorough analysis of the contextual frames of words in minute details.

Although it is understood that various fields of descriptive, applied, computational, and cognitive linguistics can benefit from the elaborate analysis of the contextual senses, the actual process of sense analysis and interpretation is a highly complex task. In this chapter, we shall try to take help of the 'corpus-based approach' to interpret as well as to understand the contextual sense of the word with reference to corpora of actual language use. In fact, an appropriate reference to corpora can help us extract, represent, and interpret the contextualized senses of words which look so enigmatic in their decontextualized appearance.

To deal with the problem of interpretation and understanding of word sense, scholars have so far proposed two major approaches:

(a) The knowledge-based approach and
(b) The corpus-based approach

In the knowledge-based approach (KBA), it is often argued that it is always better and useful to refer to information found in structured knowledge sources (e.g., *dictionary*, *grammar*, *word book*, *thesauri*, etc.) in understanding word senses (Schütze 1998; Coleman and Kay 2000; Cuyckens and Zawada 2001). On the other hand, in the corpus-based approach (CBA), scholars are interested to rely on information of word sense retrieved from the usage of words in the corpus (Ravin and Leacock 2000; Vera 2002). We like to subscribe the 'corpus-based approach' since it helps us mark the senses of words to their distributional information and contextual frames. While distributional information refers to the frequency of distribution of senses of the words in different kinds of texts, contextual frame refers to the environment of occurrence of the words in texts. We like to extract necessary information from analysis of contexts of words used in a corpus which is enriched with diverse text types to provide necessary clues to understand the actual contextual senses. Therefore, to substantiate our arguments, we like to draw supporting evidence from the Bangla corpus of written texts developed in the TDIL project.

The present chapter is organized as follows. In Sect. 10.2, we shall discuss in brief some of the propositions about word sense as proposed by earlier scholars; in Sect. 10.3, we shall try to identify the factors and issues that are considered responsible for sense variation of words; in Sect. 10.4, we shall look into the basic nature and patterns of sense variation of words as observed in all natural languages;

in Sect. 10.5, we shall try to define the contexts and how do they work behind sense variation; in Sect. 10.6, we shall try to identify the interface that lies among the contexts; and in Sect. 10.7, we shall show how data and information collected from language corpora can be fruitfully utilized in the act of sense disambiguation of words.

10.2 Propositions About Word Meaning

The traditional views of lexical semantics have been severely criticized by some modern linguists to nullify the importance of our age-old notion about the etymological meaning of words. They have claimed that the meanings of words actually come from their contextualized usages and not from the meanings recorded in their etymological information. One of the chief exponents of this model, Bronislaw Malinowski, has argued the following:

> … the meaning of any single word is to a very high degree dependent on its context … a word without linguistic context is a mere figment and stands for nothing by itself, so in reality of a spoken living tongue, the utterance has no meaning except in the context of situation. (Malinowsky 1923: 307)

Many modern semanticists including Firth (1957), Lyons (1963), Nida (1997), Cruse (2000), Goustad (2001), and others have supported almost similar observations. The basic argument of this observation is that the actual meaning of a word does not come from its origin (i.e., etymology) or its surface structure (i.e., morphological form), but from 'the company it keeps' (Firth 1957: 21). Therefore, the only way to determine the meaning of a word is to examine its usage variations in particular contexts. They have also emphasized that both cultural and linguistic contexts should be explored with equal emphasis for this purpose if required. What we understand from this deliberation is that meanings of words when the words are detached from their contextual frames of usage are incomplete. Therefore, we need varieties of context-based information for proper semantic analysis and understanding of the meaning of words.

Interestingly, some ancient Indian grammarians have also argued to refer to the role of context in the act of understanding the actual meaning or sense of words. According to ancient Indian grammarians, words are meaningful only when these are used in sentences. The literal senses of words are possible to extend, change, and revise according to their usages, while the actual senses are possible to derive from the contextual frames of their usages (Verma and Krishnaswamy 1989: 330). It has confirmed our general observation that the senses of words are indeed associated with their syntactic, topical, prosodic, idiomatic, and similar other characteristically observable contexts (Mindt 1991: 185).

This observation leads us to argue that if a word has been separated from the context of its use which is bound by various observable environments, we shall not be able to decipher its actual sense. That means we may fail to understand what the

speakers or the writers intend to convey to their listeners or readers through the use of the word. However, the process of decipherment of actual contextual sense of words is not a trivial task, since in most of the situations it is hidden within various linguistic and extralinguistic factors relating to a linguistic event. We have captured and comprehend these factors to understand the implied meanings.

10.3 Issue of Sense Variation

When we assume that words are interlinked with several senses, and we understand that the question of sense variation of words becomes an important challenge in word sense disambiguation. In general, it is noted that most of the words of a natural language are capable to exhibit multiple senses generated by their semantic extension. For example, the English word *head*, because of the feature of semantic extension, is able to refer to various senses in the following manner:

Word : Head
POS : Noun
Senses :

top of (human) body	top of glass
top of a department	number of persons
number of cattle	knowledge
intelligence	inborn ability
the top part of a thing	front part of a thing
president of a country	leader of group, team, gang
father of family	source of a river
blade of a spear	mouth of a pimple
the side of coin	edge of a bed, etc.

Similarly, the Bangla word *kathā*, due to variation in usage in different contextualized frames, is able to refer to various senses such as the followings:

Word : kathā
POS : Noun
Senses :

word	statement
description	narration
assurance	story
fact	event
opinion	promise
oath	excuse
conversation	suggestion
provocation	commitment
prescription	compulsion

context	explanation
order	request, etc.

When we try to scrutinize the senses listed above, the phenomenon of sense variation reveals a unique network of intricate relations among the senses, which hardly come out to surface for dissolution in a simple way. After analyzing the senses, we identify the following notable features of the phenomenon the understanding of which may provide clues for designing systems for word sense disambiguation.

(1) A word can have a core sense (explicit or implicit) of its own.
(2) The sense of a word may change due to the context of its use.
(3) A new sense is a conceptual extension of the core sense.
(4) Linguistic and extralinguistic factors are responsible for sense variation.
(5) If the context is understood, word sense can be captured.

The most crucial question in this regard is: How does sense vary due to variation in context? It is an important question not only is word sense understanding and sense disambiguation but also in many other domains of linguistics such as lexical semantics, lexicography, language teaching, and language cognition. Probably, information obtained from the contextual frames can provide us important clues to find a suitable answer. For this, we need to start a search for the contexts to understand what kinds of information they supply and how do they supply to us.

The availability of a large number of polysemous words (words with multiple senses) in a natural language raises a pertinent question: why some words are polysemous while others are not (Dash and Chaudhuri 2002)? It is not easy to find out strong reasons to produce a fitting answer to this question. However, we may try to identify some linguistic and extralinguistic factors that may be behind this particular phenomenon. The factors that trigger sense variations of words may be summarized in the following manners. Among the linguistic factors, the followings may be the most important ones:

(1) The context of occurrence of words in a piece of text is the biggest source of sense variation of words. That means a variation of context has a possibility of generating a new sense. Most of the polysemous words found in a natural language are of this type. In fact, context causes sense variation in so many ways that it becomes almost impossible to understand the actual sense of the words without proper reference to the contexts of their occurrence.
(2) Collocation also helps words to generate new or different senses. One can note a semantic shift when a particular word (say, target word (TW)) collocates with a neighboring word (NW) to generate a new sense. For instance, in Bangla, the word *mukh(a)* (TW) can generate various senses when it collocates with its neighboring words, as the following examples show.

mukh ālgā ‘long tongue’	mukh cchaṭā ‘glamor’
mukh bandha ‘introduction’	mukh chorā ‘bashful’
mukh chun ‘abashed look’	mukh jhāmṭā ‘scolding’
mukh khārāp ‘filthy mouth’	mukh nāṛā ‘mouthing’

mukh patra ‘manifesto’	mukh pātra ‘spokesperson’
mukh phoṛ ‘outspoken’	mukh sarbasva ‘gasbag’
mukh poṛā ‘scandalous person’	mukh rocak ‘tasteful’
mukh śrī ‘beauty’	mukh bhaṅgi ‘grimace’
mukh chabi ‘beauty’	mukh chandra ‘moon face’
mukh chāpā ‘tongue tied’	mukh kamal ‘lotus face’
mukh maṇḍal ‘face’	mukh miṣṭi ‘sweet language’
mukh padma ‘lotus’	mukh pāt ‘inauguration’
mukh ruchi ‘taste’	mukh śuddhi ‘deodorization’

In each case, the core sense of the TW is changed due to collocation with the NW. The NW performs the role of a variable to generate a new sense of the TW. It is, however, difficult to understand the new sense of the TW if we do not analyze and associate the meaning of the NW with that of the TW. This is a common feature noted in descriptive, adverbial, and reciprocal compound words used in the Bangla texts. Perhaps, it is also true to other natural languages, which maintain genealogical relation with Bangla.

(3) The change of the part-of-speech of words can be another factor behind the phenomenon of sense variation of words (Dash and Chaudhuri 2002) . It is noted that it may cause words to generate new senses, which may be different from the primary sense of the words. Although it keeps a silent relation, it often maintains a safe distance. Such conceptual expansion may add an extra shade to the actual sense of a word. For example, in Bangla, the word *chāṛā* is usually used to indicate ‘without’ which is an adverb (ADV) derived from the verb root (FV-RT) $\surd$chāṛ meaning ‘to make free something.’ The word is also found to be used as a noun (NN) to mean ‘a mature female calf which is freed from its mother’ as well as an adjective (ADJ) to refer to ‘something, which is freed.’ If all these sense variations are combined together, we can trace a relational interface existing among the senses used in different parts-of-speech of the word. However, in each part-of-speech, the word carries its own core sense (i.e., sense of separation) originally found in the root of the word.

The information gathered from various extralinguistic sources may help in understanding the nature of sense variation of words. These extralinguistic factors are not always clearly traceable within the immediate context of word use, as they originate from different sources which may have an indirect link with the words under analysis. In most cases, the majority of the factors may come from various ethnographic (e.g., *historical*, *social*, *cultural*, *moral*, *geographical*, *demographic*, *political*, *cognitive,* etc.) sources which are not often clearly understood by the users of a language. The question is how and why the information available in various extralinguistic sources becomes necessary in the study of sense variation of words.

Scholars have attempted to provide an explanation, which seems to be quite reasonable. In an extensive study on Japanese taste terms, Backhouse (1994) has pointed out the followings:

> … language is used in the world, and lexical items relate to aspects of this world: in particular, lexical items are applied to extralingual categories of entities, qualities, actions, events, and

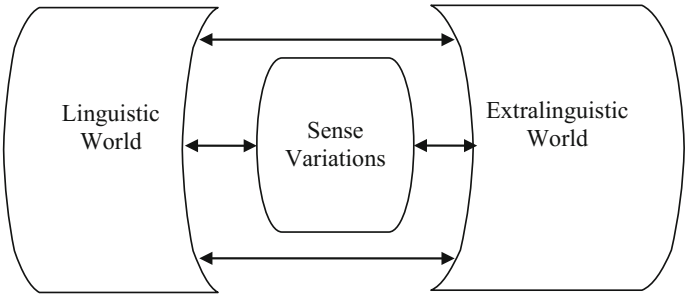

Fig. 10.1 World of information in sense variation of words

> states, and the relation between an item and such categories … is normally understood as constituting a central part of its meaning. (Backhouse 1994: 23)

Similarly, Deane (1988) has tried to justify by arguing that the human thinking process has two vital complementary characteristics. While its displays flexible responses to novel situations, it is also highly structured, incorporating detailed information about the world (Deane 1988: 325). The value of extralinguistic source is also acknowledged by Kay (2000) as he argues that since natural language is concerned with organizing our perceptions of the world we live in, it is not, therefore, unreasonable that an appeal to knowledge of the world should form a part of the process of defining lexical items we come across in life (Kay 2000: 63). Thus, scholars have tried to justify the significant of extralinguistic factors in the act of understanding sense variation of words.

In light of the arguments stated above, we can try to understand why the issue of sense variation of words is increasingly turning its attention toward the external world. What we understand is that in the act of understanding sense variation of words, we invariably require all kinds of information to probe into the phenomenon. We may seek help from the internal world of linguistics or from the extralinguistic world. If required, we may cite references from various domains of human knowledge. Information may come from every field of human knowledge for elucidating the feature of sense variation of words in question. This is a two-way process: (a) the extralinguistic world relating to life and living of speakers shades light on sense variation and (b) the process of elucidation of senses of words shades new light on life and society of the speakers (Fig. 10.1).

10.4 Nature of Sense Variation

One interesting finding in this context is that not only the function words, the content words are also quite dynamic in sense variations. We find that in the Bangla text corpus words belonging to the class of noun, verb, and adjective are very dynamic in sense variation as they frequently change senses based on different contexts and

Table 10.1 List of Bangla nouns with 15+ sense variations

Noun	Core sense	Senses	Noun	Core sense	Senses
uttar	North	15+	kathā	word	25+
kāj	work	15+	kāl	time	20+
kul	shore	20+	kuṭ	mountain	20+
koṣ	cell	15+	kṣetra	field	15+
gati	speed	18+	gun	quality	20+
guru	master	21+	ghar	home	15+
cāl	gait	19+	daśā	stage	15+
dharma	religion	20+	nām	name	17+
pakṣa	side	15+	pad	position	17+
parba	phase	16+	pātā	leaf	16+
bhāb	mood	25+	māthā	head	30+
mukh	face	16+	ýug	era	23+
ras	ras	27+	samay	time	15+
sthān	place	20+	hāt	hand	15+

Table 10.2 Some Bangla verbs with 20+ sense variations

Verb	Core sense	Senses	Verb	Core sense	Senses
āsā	To come	21+	oṭhā	to rise	27+
karā	To do	45+	kāṭā	to cut	23+
khāoyā	To eat	100+	calā	to walk	20+
chāṛā	To free	22+	tolā	to lift	30+
thākā	To stay	21+	deoyā	to give	33+
dekhā	To see	25+	dharā	to catch	40+
paṛā	To read	27+	phoṭā	to bloom	30+
basā	To sit	22+	bharā	to fill	25+
mārā	To kill	22+	rākhā	to keep	24+
lāgā	To stick	26+	saoyā	to bear	20+

situations. Moreover, many postpositions also show sense variation. In sum, most of the words belonging to this parts-of-speech denote multiple senses in various contexts without changing their orthographic forms. These words are normally included in a dictionary either as separate entries or as semantic extensions of a single entry. Such words are not very large in number. Also, searching in a Bangla text corpus we have collected a large number of words (both function and content words), which are mostly polysemous. To support our observation, we have presented below some nouns (Table 10.1), verbs (Table 10.2), and adjectives (Table 10.3), which exhibit are polysemous in nature.

Table 10.3 Some Bangla adjectives with 15+ sense variations

Adjective	Core sense	Senses	Adjective	Core sense	Senses
kā̃cā	Raw	25+	pākā	ripe	25+
khara	Rude	17+	khārāp	ugly	25+
khāli	Empty	15+	garam	hot	21+
ghana	Thick	15+	kālo	black	18+
caram	Ultimate	15+	choṭa	small	15+
naṣṭa	Spoiled	17+	naram	soft	15+
baṛa	big	20+	manda	bad	20+
miṣṭi	sweet	15+	mukta	free	16+
mṛdu	light	17+	laghu	light	20+
śakta	hard	20+	sādā	white	15+

In case of verbs (Table 10.2), the most interesting point to note is that the majority of these verbs are also available to be used as adjectives (as participial forms derived from verbs) and nouns (as gerunds or verbal nouns) in the language. In both the lexical categories, they preserve their polysemous nature. The number of sense variations in their respective lexical categories is, however, not equal to that of their original verb category. That means from a verb like *khāoyā* 'to eat,' we can have 100+ senses including all its use as a verb, adjective, and noun. Furthermore, a particular sense denoted by the word as a verb may also encompass the range of senses denoted by its adjectival and nominal forms. Therefore, the number of sense variation of the word, in other lexical categories, is not always equal.

A simple comparison of the words included in the lists (Tables 10.1, 10.2, and 10.3) can reveal that most of the words are easy in form and quite frequent in use in the language. This confirms our hypothesis that most of the common words are mostly polysemous in nature due to frequent use in different contexts. The interesting thing is that the number of sense variation of verbs is higher than that of nouns and adjectives. It is not clear why verbs can denote more senses than nouns and adjectives. Also, there are a few postpositions which denote comparatively fewer sense variations (2–8). Since we cannot present all the words with their total range of sense variation with reference to their contextual usages, we present below a Bangla noun with its possible list of senses obtained from a Bangla text corpus by applying a concordance program (Table 10.4).

Word : māthā
Lexical class : Noun
Core meaning : 'head'
No. of senses : 30+

Table 10.4 Sense variation of Bangla word *māthā* in contexts

Word	Gloss	Word	Gloss
mānuṣer māthā	human head	gācher māthā	treetop
chātār māthā	useless thing	ṭebiler māthā	tabletop
grāmer māthā	the village head	pāhāṛer māthā	the mountain peak
āṅguler māthā	fingertip	jaler māthā	water surface
nadīr māthā	the river mouth	byaṅger māthā	nonsense
rāstār māthā	road end	kā̃cā māthā	tender mind
pariskār māthā	sharp brain	kṣurer māthā	blade of razor
kājer māthā	work sense	aṅker māthā	knowledge in math
daler māthā	team leader	gyaṅger māthā	gang master
naukār māthā	head of a boat	aphiser māthā	office boss
kompānir māthā	owner of a firm	paribārer māthā	head of the family
pākā māthā	expert head	kāgajer māthā	the margin of the paper
bichānār māthā	edge of the bed	rāṣṭrer māthā	president of the country
phoṛār māthā	pimple mouth	mudrār māthā	the side of the coin
pener māthā	the cap of the pen	sojā māthā	straight head
ũcu māthā	high pride	māthā muṇḍu	head and tail

10.5 Context in Sense Variation

Context plays utmost importance in word sense understanding. It not only triggers variation in sense but also supplies information to know why and how words vary in sense. Due to this factor, the context has been an issue of high importance in semantics, language technology, cognitive linguistics, lexicography, and language teaching (Dash 2008). Keeping this in the background, we shall make attempt here to explore nature, type, and role of context in word sense understanding. In the course of our discussion, we shall try to identify different types of context that play roles in triggering sense variation of words. We shall argue that a corpus is a useful resource that provides necessary information to identify all types of context and to understand their role in sense variation. At certain situations, a reference to a particular context may be useful; but the reference to other contexts may be useful for deciphering senses embedded within words in a text.

A word, when it occurs in a particular context, usually denotes only one sense out of multiple possible senses. How it happens is an enigma. The assumption is that it is the context that discriminates among senses to determine only one sense. If this assumption is taken for granted, we need a method of sense capturing that can determine context, since there is no fixed system by which we can automatically identify the context. It is believed that identification of context depends heavily on

the intuitive ability of language users. This brings us nowhere near the problem. Rather it leads us to believe that people with richer intuition can excel over others in this task. This cannot be the right way for solving a long-standing problem.

A properly sampled language corpus is considered quite useful in this case as most of the words show multifaceted representation in a corpus with adequate contextual evidence. Moreover, special corpus with exclusive examples of the contextualized use of words is used as a supporting database (Dash 2005a, b: 9). Before we explore how a corpus contributes to the act of context search, let us first understand what a context means and what its basic properties are.

In linguistics, *context* refers to an immediate environment in which a linguistic item occurs. It can be a phoneme, a morpheme, a word, a phrase, a clause, a sentence. For the present study, we consider the context of word use only. Not necessarily, the context of words is explicit in all situations. Sometimes, it is hidden within neighboring areas of use of a word or may be located at distant places linked to a text or a topic. It implies that we cannot always extract relevant information of use of words from the immediate environment. We have to take the topic of discussion under consideration because necessary information may be hidden here. Taking these issues into account, Miller and Leacock (2000) divided context into two broad types:

(a) Local Context (LC): One/two words before and after a word the sense which we are interested in. It is target word (TW).
(b) Topical Context (TC). The discourse and topic of a text where the TW is used.

In our view, the two contexts are not enough as they fail to provide necessary and sufficient information we need to understand the intended sense of a word. At certain readings, information available in the two contexts is sufficient, but these are not the ultimate source of all possible readings. Since we need more information to understand the actual sense of a word used in a text, we classify context into four major types:

(a) Local Context (LC): TW with one word before and after it.
(b) Sentential Context (SC): Sentence where TW occurs.
(c) Topical Context (TC): Topic or content of a text.
(d) Global Context (GC): Extralinguistic information.

The conceptual hierarchical layering of contexts may be understood from the following diagram (Fig. 10.2).

In this diagram, LC is made with a TW along with one word before and after it. The LC occupies the center of all attention as it provides the basic information about sense variations of a word. Therefore, LC is accessed first to obtain information from neighboring words of the TW. If this is not enough, we refer to SC to retrieve information from the sentence where TW occurs. Next, we access TC to acquire further information from content or topic of a text. Finally, we access GC to gather information from the extralinguistic world (i.e., *discourse*, *pragmatics*, *world knowledge*, etc.). The process that one can adopt for extracting information from all contexts is shown in the following diagram which also displays the role of contexts in the generation of new senses of words (Fig. 10.3).

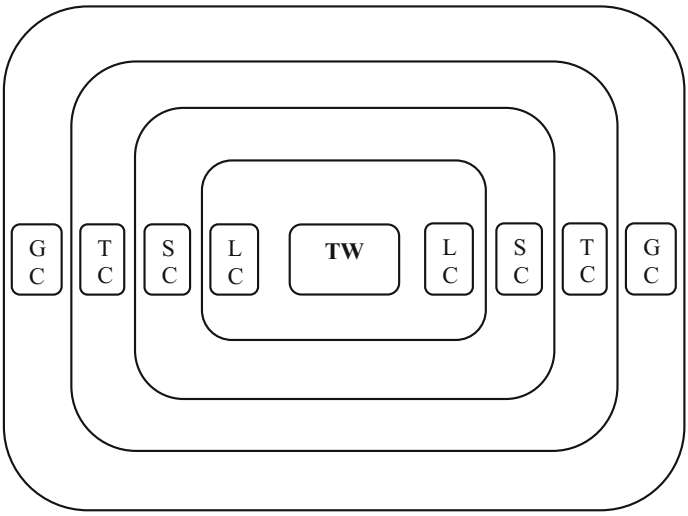

Fig. 10.2 Conceptual hierarchical layers of contexts

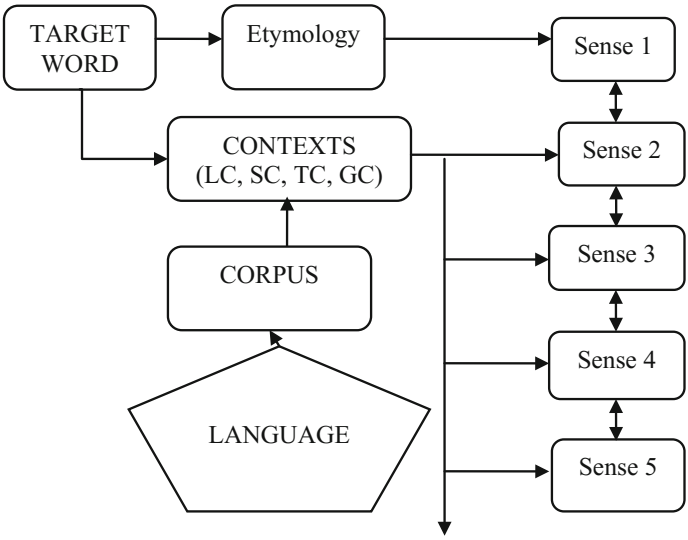

Fig. 10.3 Generation of new sense due to the variation of contexts

10.6 Interfaces Among the Contexts

Sense variation of words is generated due to use of information from various contexts. This may create an impression that each context is characteristically different from other; that they do not have any link; and that they do not maintain conceptual relation

among themselves. It may also give an impression that one has to move in a sequential order to reach to the final context.

In reality, each context is interlinked with other through an invisible thread of interdependency. So we have freedom in use of information in one context while dealing with other. Also, we can use information from all contexts together to solve the problem of sense variation. As there is no question for sequential use of contextual information, we can start with any context and move to other as the situation demands. For instance, if we require information from GC while dealing with LC, we can do that. To understand how the interface works, consider the example given below.

(1) The public has well relished your Sunday soup in the tabloid.

To understand the sense of the TW (*relished*), one needs to explore information from all contexts since the TW is used here in a discrete manner with a metaphoric sense. To know the actual sense of the TW, one must have answers to the following questions:

(a) Who has made the statement?
(b) To whom the statement is made?
(c) When is it said?
(d) Where is it said?
(e) What does the word *relish* mean?
(f) What does the phrase '*relished your Sunday soup*' refer to?
(g) How does '*Sunday soup*' become relishing to the public?

A reading of the word string [i.e., *relished…soup*] shows that it is used in a figurative sense. Information collected from all the sources help to extract an inner sense of the construction, conceive interface inherent within the network of time–place–agent–action, and capture the actual contextual sense of the word.

The LC carries primary importance in understanding sense variation of words. We can access SC, TC, and GC only when we are not happy with information from LC. Reference to other contexts comes in the subsequent stages when information obtained from LC is not enough. The SC, which refers to a sentence where the TW occurs, may include the immediate environment of TW focusing on both of its neighboring words and distal words. The TC refers to topic or content of a text where TW occurs. At the cognitive level, it tries to fabricate a sense relation between TW and topic of a text. Finally, through GC, we try to refer to the extralinguistic domain where from we gather all kinds of information relating to the external world (Dash 2005a, b).

The degree of understanding of sense variation of words depends on width and depth of world knowledge of a user. People with greater linguistic and extralinguistic knowledge and experience are more efficient in understanding the contextual sense of words than others. For instance, a native English speaker with better experience of English life, language, society, and culture can easily catch the inner sense of *soup* from the sentence (1) than those who have limited knowledge of English life and living, ideas and language, and society and culture.

10.7 Corpus in Sense Disambiguation

A corpus is a reliable source of information for identifying the wider range of senses a word denotes as well as extracting actual contextual sense (Dash 2008). A corpus makes a significant contribution toward empirical analysis of contexts with close reference to usage. Standard dictionaries usually fail here because the range of sense variation of a word that shows up in a corpus exceeds the number of senses listed in a dictionary (Fillmore and Atkins 2000).

The availability of corpus simplifies the process of accessing contexts of use of words. One can retrieve all uses of a word from a corpus in the form of a concordance to analyze with supporting instances the range of sense variation, patterns of sense change, nature of sense variation, factors behind sense variation, etc. (Kilgarriff 2001). Thus, a corpus can save us from referring to linguistic intuition for obtaining possible senses of a word. A corpus supplies much more information than that available from our linguistic knowledge and native language intuition.

A corpus directs us to locate all contexts systematically, classifies them based on their features, and accesses them for necessary information. In fact, a corpus makes the task of sense understanding much easier which is not possible through intuitive evidence (Gale et al. 1992). A corpus enriched with various types of word use adds an extra shade to linguistics so far unknown in intuitive frames. It excels over intuition because it supplies a wider spectrum of contexts of word use as well as provides necessary contextual clues for understanding variation in senses. Evidence gathered from corpus shows that some linguistic issues (e.g., *structure*, *part-of-speech*, *synonymy*, *lexical collocation*, *lexical gap*, *usage*, *co-text*, etc.) control the event of sense variation of words. Also, extralinguistic factors (e.g., *figurative usage*, *idiomatic usage*, *metaphors*, *pragmatics*, *discourse*, *dialogic interactions*, *sociocultural settings*, *register variables*, etc.) contribute to the process of sense variation of words.

Variation of sense of a word may come from various sources: internal morphemic structure, etymological history, dictionary identity, lexical class, grammatical function, synonyms, antonyms, contextual occurrence, lexical association, lexical collocation, usage, content of discourse, and similar other factors (Ide and Véronis 1998; Kilgarriff and Palmer 2000). A corpus is able to refer to all possible contexts for understanding senses. It supplies all linguistic (e.g., *morphological*, *lexical*, *grammatical*, *semantic*, *syntactic*, etc.) and extralinguistic (e.g., *telic*, *deictic*, *temporal*, *tense*, *spatial*, *event*, *pragmatics and discourse*, etc.) information essential in identifying sense variation, observing total sense range, differentiating among related senses, and extracting contextual sense. Once we analyze relevant information as well as capture all possible sense variations, we succeed to understand the target sense. The following diagram shows how a corpus may be used to know the contextual sense of a word (Fig. 10.4).

In a corpus, we come across a large number of words, which may or may not have sense variation. Our first task is to find out if a word is denoting sense variation. To solve this, we can use information from two sources: dictionary, and corpus.

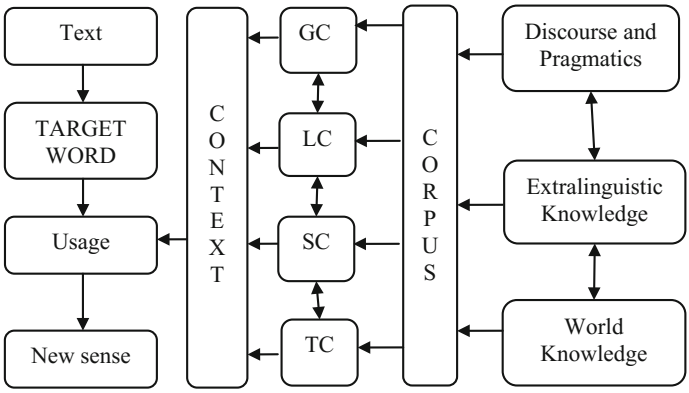

Fig. 10.4 Access corpus to know the actual contextual sense of word

Reference to dictionary supplies clues to know if a word has sense variation as multiple subentries in meaning in a dictionary determines the polysemous identity of a word. Reference to a corpus gives information about the range and patterns of sense variation. Possibly, the total range of sense variation may be noted if a corpus contains large representative samples from all text types with wider genre and subject variations.

Once a word is found to have sense variation, we need to know the two things:

(a) How many sense variations does it possess?
(b) In which (contextual) sense it is used?

We can refer to the context-free and context-bound information to find answers to the questions. Context-free information is obtained from structured knowledge sources (*grammar*, *morphology*, *dictionary*, *thesaurus*, *etymology*, etc.). Context-bound information is obtained from a corpus with reference to LC, SC, TC, and GC. It is not likely that information obtained from all sources is used whenever one tries to understand sense variation. But it helps to understand the feature of sense variation and obtain actual contextual sense from the score of multiple senses. Both the process can work in tandem. For this, we propose the Access of Information from Multiple Sources (AIMS) method for extracting both types of information (Fig. 10.5). By using a step-by-step process of 'input–analysis–output' scheme, one can get necessary information from intralinguistic and extralinguistic contexts to understand sense variations as well as to extract actual sense.

We suggest for using information from various subfields of words in the context-free situation. Information obtained from suffix and case endings usually plays a decisive role in sense disambiguation. This is useful because many inflected words are sidelined to a particular part-of-speech and sense after using class-specific suffix or case ending. For example, the *call* is a noun or a verb with two different senses. It is a noun if an agentive case marker is tagged to it (e.g., call + $\text{-er}_{[\text{Case\_Agent}]}$ > $\text{caller}_{[\text{N}]}$); it is a verb if a verb suffix is tagged (e.g., call + $\text{-ed}_{[\text{FV\_PST}]}$ > $\text{called}_{[\text{FV}]}$). Thus, iden-

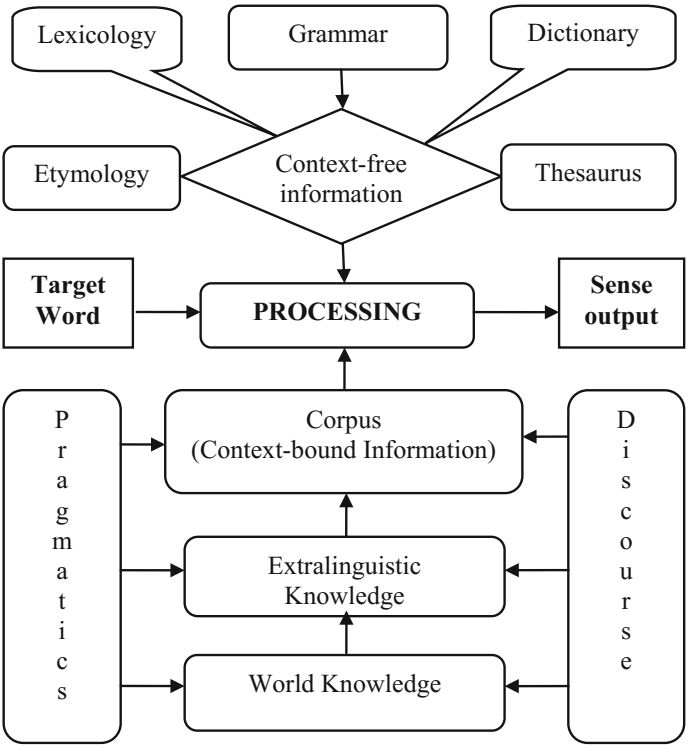

Fig. 10.5 AIMS method for word sense disambiguation

tification of suffix marker or case ending of a word helps to put them to a particular lexical class and sense without further reference to the context of its use. In the same manner, based on need, we refer to etymological, lexical, dictionary, and thesaural information of a word to understand its sense variation in the context-free situation (Dash 2008).

However, if we find that information obtained from subfields relating to the context-free situation is not sufficient, we can refer to the context-bound information found in a corpus, and information retrieved from extralinguistic sources. In essence, the AIMS method has the liberty to refer to all kinds of information. It refers to lexical information stored in a dictionary, morphophonemic information found in grammars, contextual information found in the corpus, figurative information found in usage, and extralinguistic information found in the external world. The AIMS method thus becomes useful in understanding sense variation of words. In AIMS, one applies linguistic information acquired from structured knowledge sources and non-linguistic information obtained from the corpus, discourse, and the world at large.

10.8 Conclusion

A word carries information of phonetics, phonology, morphology, morphophonemics, lexicology, semantics, syntax, morphosyntax, text, grammar, co-text, etymology, metaphor, discourse, pragmatics, world knowledge, and others (Pinker 1995: 344). It is not easy to find all information just by looking at its surface form. One needs a good system along with strong linguistic intuition to understand all explicit and implicit senses a word denotes in a language.

It is not necessary to define all possible senses of a word (Moravcsik 2001). If we do this, we severely damage productivity and flexibility of a language, overload lexicon, and burden language learners. Sense variation is a vital aspect of a natural language. It leaves many things in a state of incompleteness out of which lexical productive strategies generate literal or metaphoric alternatives to accommodate novel experiences and situations.

In language processing, lexicography, translation, and many other fields of language technology, we face the problem of sense variation. We want to understand the phenomenon in details so that we can apply our knowledge in sense discrimination, information retrieval, content analysis, WordNet, language cognition, text alignment, parsing, machine learning, etc. Also, we need elaborate information of sense variation for compiling a dictionary, linguistic theory making, and teaching a language. Finally, systematic study of sense variation of words helps us understand 'semantic indeterminacy' and 'sense gradiance' of words in the act of language cognition.

References

Backhouse, A.E. 1994. *The Lexical Field of Taste: A Semantic Study of Japanese Taste Terms*. Cambridge: Cambridge University Press.

Coleman, J., and C.J. Kay (eds.). 2000. *Lexicology, Semantics, and Lexicography*. Amsterdam-Philadelphia: John Benjamins.

Cruse, A. 2000. *Meaning in Language: An Introduction to Semantics and Pragmatics*. Oxford: Oxford University Press.

Cuyckens, H., and B. Zawada (eds.). 2001. *Polysemy in Cognitive Linguistics*. Amsterdam/Philadelphia: John Benjamins.

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

Dash, N.S. 2005b. The Role of Context in Sense Variation: Introducing Corpus Linguistics in Indian Contexts. *Language in India* 5 (6): 12–32.

Dash, N.S. 2008. Context and Contextual Word Meaning. *SKASE Journal of Theoretical Linguistics* 5 (2): 21–31.

Dash, N.S., and B.B. Chaudhuri. 2002. Using Text Corpora for Understanding Polysemy in Bangla. In *Proceedings of IEEE Language Engineering Conference*, Department of Computer Science and Engineering, Central University, Hyderabad, 13–15 November 2002, 99–109.

Deane, P.D. 1988. Polysemy and Cognition. *Lingua* 75: 325–361.

Fillmore, C.J., and B.T.S. Atkins. 2000. Describing Polysemy: The Case of 'Crawl'. In *Polysemy: Theoretical and Computational Approaches*, ed. Y. Ravin and C. Leacock, 91–110. New York: Oxford University Press Inc.

Firth, J.R. 1957. Modes of Meaning. In *Papers in Linguistics 1934-1951*. London: Oxford University Press.
Gale, W., K.W. Church, and D. Yarowsky. 1992. A method for disambiguating word senses in a large corpus. *Computers and the Humanities* 26 (4): 415–439.
Goustad, T. 2001. Statistical Corpus-Based Word Sense Disambiguation: Pseudo-Words Vs. Real Ambiguous Words. In *Companion Volume to the Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics* (ACL-2001), 24–29.
Ide, N., and J. Véronis (eds.). 1998. *Computational Linguistics. Special Issue on Word Sense Disambiguation*, vol. 24, no. 1.
Kay, C.J. 2000. Historical Semantics and Historical Lexicography: will the Twin ever Meet? In *Lexicology, Semantics, and Lexicography*, ed. J. Coleman and C.J. Kay, 53–68. Amsterdam-Philadelphia: John Benjamins.
Kilgarriff, A. 2001. Generative Lexicon Meets Corpus Data: The Case of Non-Standard Word Uses. In *The Language of Word Meaning*, ed. P. Bouillon and F. Busa, 312–328. Cambridge: Cambridge University Press.
Kilgariff, A., and J. Palmer (eds.). 2000. *Computer and the Humanities: Special Issue on Word Sense Disambiguation*, vol. 34, no. 1.
Lyons, J. 1963. *Structural Semantics*. Cambridge: Cambridge University Press.
Malinowsky, B. 1923. The Problem of Meaning in Primitive Languages. In: *The Meaning of Meaning*, ed. C.K. Ogden, I.A. Richards, 52–65, 9th ed. London: Keegan and Paul.
Miller, G.A., and C. Leacock. 2000. Lexical Representations for Sentence Processing. In *Polysemy: Theoretical and Computational Approaches*, ed. Y. Ravin and C. Leacock, 151–160. New York: Oxford University Press Inc.
Mindt, D. 1991. Syntactic Evidence for Semantic Distinctions in English. In *English Corpus Linguistics. Studies in Honour of Jan Svartvik*, ed. by K. Aijmer, B. Altenberg, 182–196. London: Longman.
Moravcsik, J.M. 2001. Metaphor; Creative Understanding and the Generative Lexicon. In *The Language of Word Meaning*, ed. P. Bouillon and F. Busa, 247–261. Cambridge: Cambridge University Press.
Nida, E.A. 1997. The Molecular Level of Lexical Semantics. *International Journal of Lexicography* 10 (4): 265–274.
Pinker, S. 1995. *The Language Instinct: The New Science of Language and Mind*. England: Penguin Books.
Ravin, Y., and C. Leacock (eds.). 2000. *Polysemy: Theoretical and Computational Approaches*. New York: Oxford University Press Inc.
Schütze, H. 1998. Automatic Word Sense Disambiguation. *Computational Linguistics* 24 (1): 97–123.
Vera, D.E.J. 2002. *A Changing World of Words: Studies in English Historical Lexicography, Lexicology and Semantics*. Amsterdam: Rodopi.
Verma, S.K., and N. Krishnaswamy. 1989. *Modern Linguistics: An Introduction*. Delhi: Oxford University Press.

Web Links

http://lct-master.org/files/WSD.pdf.
http://www.cs.uu.nl/docs/vakken/ll/word_sense_disamb.pdf.
http://aclweb.org/aclwiki/index.php?title=Word_sense_disambiguation.
http://wwwusers.di.uniroma1.it/~navigli/pubs/ACM_Survey_2009_Navigli.pdf.
http://www.scholarpedia.org/article/Word_sense_disambiguation.
http://www.cs.upc.edu/~escudero/wsd/06-tesi.pdf.

Chapter 11
Corpus and Technical TermBank

Abstract The development of an exhaustive database of scientific and technical terms in a natural language carries tremendous importance in the areas of linguistic resource generation, translation, machine learning, knowledge representation, language standardization, information retrieval, dictionary compilation, language education, text composition, language planning, as well as in many other domains of language technology and mass literacy. Keeping these utilities in mind, in this chapter, we propose for developing a large lexical database of scientific and technical terms in a natural language with the utilization of a corpus. In this work, we propose to adopt an advanced method for systematic collection of scientific and technical terms from a digital language corpus, which is already developed and made available for general access. Since most of the Indian languages are enriched with digital texts that are available on the Internet, it will not be unfair to expect that we can develop a resource of this kind in most of the Indian languages. Following some of the stages of corpus processing discussed in this book (Chap. 5), we can develop an exhaustive database of scientific and technical terms in any of the Indian languages which can be utilized in all possible linguistic activities.

Keywords Scientific term · Technical term · Processing · Corpus · Part-of-speech tagging · Concordance · Collocation · Lemmatization · Frequency sorting Type–token analysis · TermBank

11.1 Introduction

The scientific and technical terms that are found to be used in a high amount in scientific and technical texts (also sometimes in general text) have been an area of real challenge in the act of sense understanding, sense disambiguation, translation, and many other works. These lexical items are indeed strong barriers to the common readers of a text as most of the common people, due to the lack of proper technical knowledge, often fail to understand the senses implied by these terms. This appears to be true when we look at the texts produced as doctor's prescription where several

N. S. Dash and L. Ramamoorthy, *Utility and Application of Language Corpora*,
https://doi.org/10.1007/978-981-13-1801-6_11

medical terms appear to be so complex and unintelligible that patients are often confused in deciphering the senses of the terms used in the prescription. A similar situation happens when we buy medicines. Many medical terms are used in the medicine box which we cannot understand. Such situations are quite frequent in most of the Indian languages where there are no scientific and technical term databases available for us to use in understanding the terms used in technical texts (Dash 2008).

If we keep these challenges in mind, we have to agree to the proposal that it is necessary to identify the scientific and technical terms in texts, isolate them, and develop special terminology databases for general as well as specialized applications. We propose to develop these from corpus through several stages of corpus processing and access, which are discussed sequentially in this chapter. In our view, the scientific and technical terminology database developed in the manner proposed here can make a strong contribution to the improvement of linguistic resources for the Indian languages and their users.

In Sect. 11.2, we discuss in brief about the nature and characteristic features of scientific terms; in Sect. 11.3, we describe the nature and characteristic features of technical terms and how they differ from scientific terms; in Sect. 11.4, we discuss some of the methods of corpus processing that include techniques like part-of-speech tagging, concordance, collocation, lemmatization, frequency sorting, and type–token analysis which are considered absolutely necessary in the act of TermBank generation; in Sect. 11.5, we present an architecture that can be used for terminology database generation; and in Sect. 11.6, we identify the people who will be interested to use the terminology database to address different needs in linguistics, applied linguistics, language technology, and allied disciplines.

11.2 Scientific Term

The expression scientific term (ST) refers to a single as well as a multiword unit that is used in different types of scientific texts in specialized senses. The literal meaning of the expression although refers to specialized terms used in scientific texts; in reality, it is not confined to the fields of science only. Rather, it encompasses all specialized terms used in any discipline of human knowledge.

Based on the connotative meaning of the expression stated above, the term 'scientific terminology' refers to the analysis of form, function, usage, and meaning of scientific terms within any area of human knowledge. Therefore, scientific terminology, in principle, investigates, among other things, how various specialized terms have come into existence and their interrelationships within the field of study. Thus, the discipline scientific terminology refers to a more formal study, which systematically investigates labeling or designating of concepts particular to different domains of human science.

Scientific terminology also involves research and analysis of the terms with a purpose of documenting and promoting their correct usage among the common people. Since these terms are not often well accepted, rightly interpreted, and properly

understood by the common people, it is necessary to have an area where all these scientific terms are clearly defined and explained for the understanding by the common people. The study may be limited to one language or can cover more than one language at the same time to develop databases of bilingual scientific terminology or multilingual scientific terminology. Similarly, it can be extended over two or more domains to focus on studies of scientific terms across several domains or subject fields.

The importance of scientific terms is not limited to information retrieval only. It is also related to the conveyance of concepts and meanings. However, it should be kept in mind that the word 'term' (i.e., index terms) used in the context of information retrieval is not the same as 'term' used in the context of scientific terminology since the word 'term' in information retrieval does not always mean any kind of specific scientific term of a particular discipline. In case of scientific terminology, the actual value of scientific terms can be measured by their contexts of usage, careful classification of terms, and their interpretation by senses they denote when used in a particular text or subject domain. So, while investigating scientific terms, the discipline 'scientific terminology' needs to adhere to several relational issues. The analysis of concepts and concept structures of scientific terms used in particular fields or domains of activity may involve the following activities:

(1) Identification of terms assigned to concepts,
(2) Analysis of the sense of the terms used,
(3) Compilation of scientific terminology database,
(4) Management of scientific terminology databases,
(5) Creation of new terms as and when required,
(6) Establishment of conceptual correspondence between terms and senses,
(7) Mapping of scientific terms used in bilingual and multilingual texts.

When we feel the need for generating a database of scientific terms to address the needs of a particular discipline, we need to deal with several issues such as generation of new concepts, reference to new ideas and materials, identification of new techniques and devices, providing alternative meanings to common words, formation of new composite words, generation of acronyms. In fact, the issues and conditions that we often apply in neologism are equally applicable for the formation of scientific terms as it asks for the consideration of several linguistic and functional issues, which we cannot have the liberty to ignore at the time of coining new scientific terms. For understanding scientific terms, the remark made by Barnhart (1978) appears quite sensible:

> The vocabulary of science should be related to the general vocabulary of educated people so that the particular contributions of any science to our knowledge and understanding of the universe can be made a part of general knowledge. The basic terms of scientific and technical vocabulary should be so explained that the beginning student can comprehend them and relate them to his experience. It should be possible, both in general purpose dictionaries and in specialized technical dictionaries to show that scientific terms are not merely hard words

> but results of a different and more exact structuring of the world by the scientist; parallel defining is of great importance as a cross-reference to closely related terms. The concept of the atom is related to molecule and nucleus and proton; one term cannot really be understood without the others. (Barnhart 1978: 1927)

From the perspective of the application, a database of scientific terms that have been developed from a language with due importance on form, function, usage, and meaning of the terms, can be used in manual and machine translation, in teaching usage patterns of scientific terms in academic and translation schools, and in the composition of scientific texts. For example, students, while studying linguistics, may come across many new terms which are full of new ideas and concepts. They need to learn these new terms not only for enriching their knowledge about the subject but also for expanding their comprehension skill in the subject. Although most of these terms are understood by seasoned linguists, they often remain as hurdles for the newcomers as well as common users. It is, therefore, necessary to make these terms understandable to the common people so that these terms gradually become a part of the common vocabulary of a language.

11.3 Technical Term

The colloquial use of 'terms' should not be confused with technical terms (TT) as these are different in sense denotation. Technical terms are used to define ideas and concepts within a special discipline or a field of specialty. The 'technical terminology,' therefore, is a highly specialized vocabulary or the nomenclature of particular disciplines. These terms have specific definitions within the field, which is not necessarily the same as their meaning in common use. We can perhaps use the term 'jargon,' which is nearest in sense; but the term 'jargon' is more informal in definition and use, while technical terms have meanings strictly defined by the disciplines. By a simple definition, a 'technical term' is a unique lexical item which has a truly specialized meaning within a specific field of human science. It implies that a word or a phrase is highly typical within a particular field of study, and only the people directly linked to this field are familiar with this and use this.

Since technical terms exist in the scale of a continuum of formality, their short definitions are formally recognized, documented, and taught by the experts working in the field while 'formal terms' are more colloquial and used by the practitioners of the field. The boundaries between 'formal terms' and 'technical terms' are, however, quite fluid, as they slide in and out of recognition quite rapidly. For instance, in the rapidly changing world of computer technology, the term 'firewall' (in the sense of a device used to filter network traffic chaos) was at first a technical term. As this device becomes more and more user-oriented, the term is widely understood and is adopted in formal terminology.

Usually, a technical term evolves as a result of needs of the experts working in a particular discipline to communicate with high precision and brevity. It often has an effect of excluding the people who are not familiar with the specialized language of

a particular discipline. This can create severe difficulties for the common people, for example, when a patient fails to follow the discussions of the medical practitioners, and thus cannot understand his own condition and treatment.

A technical term should have the qualities to be scientifically accurate and intelligible to the people of the discipline. As we cannot do justice with the definition of all technical terms in general, we can take help from the experts of different fields for understanding these terms. It is better to have a board of experts in different fields who can advise us in the matter of defining the technical terms used in different disciplines.

Since all the technical terms cannot find a place in a general reference dictionary, it becomes mandatory to select those technical terms which are considered eligible for inclusion in a reference dictionary. However, in case of a general database of technical terminology, each and every technical term is an automatic choice where there is no option open for preference or rejection of any technical term or the other. The newly coined technical terms, as well as the old technical terms attaching new meanings, are eligible for inclusion in a general database of technical terminology.

The problem relating to commonness and uncommonness of technical terms is an important issue, which can be addressed with consideration of goals of a general database of technical terms. Even if some technical terms appear artificial and ambiguous, these should be included in a general technical terminology database and should be given an equal amount of importance as done for the common ones. Moreover, in those cases where we find several technical terms denoting one concept or one term denoting several concepts we have to be very careful with regard to their inclusion in the term database. All these are equally eligible for inclusion in a general technical terminology database, but proper attention has to be paid for their specific sense denotation in different contexts.

With regard to the definition of technical terms, we argue that the definition of a technical term should be provided in such a manner that it becomes clearly intelligible to the non-specialists as well. This argument can settle the questions whether only a technical definition should be given or a simple general definition should also be there. For elucidation, let us consider the following examples (Table 11.1).

The examples given above show that the definitions of technical terms should be formed in such manner that these are able to transmit the ideas of science and technology into general language for an understanding of the common people (Table 11.1). If it is not possible to provide a precise definition of the terms, these can be explained with some equivalent illustrative terms and pictorials so that the illustration becomes far more expressive for understanding the items or objects by the common mass. By the way of opting visual illustration, we are actually entering into a domain of encyclopedic definition, which is necessary for most of the technical terms of science, engineering, and technology.

The availability of databases of scientific and technical terms in most of the advanced languages like English, German, French, Spanish, Chinese, and Japanese is a lesson for the less resourced languages. The way these resources are developed and utilized for these languages can be adopted for the same purposes for the less advanced as well as less resourced languages. In this context, it is quite painful to

Table 11.1 Difference between a technical definition and a formal definition

Term	Technical definition	Formal definition
Iron	It is the second heaviest stable isotope produced by the alpha process in stellar nucleosynthesis, made with a chemical element with symbol Fe (Ferrum) and atomic number 26. It is a group 8 and a period 4 element	It is a metal with lustrous and silvery color. It is the most abundant element in the core of meteorites and in the dense metal cores of planets such as earth. It is one of the most common sources of ferromagnetic materials adopted in everyday use
Coal	It is a fossil fuel formed in an ecosystem where plant remains were preserved by water and mud through oxidization and biodegradation, and its chemical and physical properties have been changed as a result of geological action over time, thus sequestering atmospheric carbon. It is a readily combustible black or brownish-black rock, composed primarily of carbon and hydrogen along with small quantities of other elements, notably sulfur	It is a hard opaque black or blackish mineral or vegetable matter found in seams or strata below the surface of the earth and used as a fuel and in the manufacture of gas, tar, etc.

record that the development of a database of scientific and technical terms for most of the Indian languages is still a far cry, although at least 10 years ago, we have developed workable digital corpora for most of the Indian national languages (Dash 2007a, b, c, d, e).

In the present context of generating scientific and technical terminology database for the Indian national languages, it is necessary to divert our attention toward this area and engage a large team of experts in this task for the benefit of the Indian languages and their speakers. The first work that we need to do in this context is to process the Indian language corpora developed so far in various ways as proposed in the following section (Dash and Basu 2012).

11.4 Processing a Language Corpus

Processing a digital language corpus is the first step toward the generation of a database of scientific and technical terms. However, the important part is that we have to identify which processing techniques are going to be maximally useful for this purpose. Also, we require efficient techniques for processing a corpus so that we are able to create an easy functional interface between theoretical and applied linguistics.

The task of corpus processing starts after the accumulation of a large amount of language data in digital form in a corpus. Once this corpus is made ready, a system developer devises appropriate techniques for processing a corpus for extracting rel-

evant linguistic data and information. There are several techniques that are used for processing a corpus; some of which are discussed here keeping their relevance in the generation of a terminology database. The corpus processing techniques that we think to be absolutely necessary for generating scientific and technical terminology database are parts-of-speech (POS) tagging, concordance, collocation, lemmatization, frequency count, lexical sorting, and type–token analysis (Dash 2005: 155). However, we have to keep all language-specific and script-related issues in mind while trying to apply these processes on a corpus to achieve our goals.

Although several corpus processing software and NLP toolkits are freely available for most of the advanced languages like English, French, German, Spanish, and others, almost nothing, except a few toy tools and systems, is available for the Indian languages. After a minute survey of the TDIL Data Centre of the Government of India, we find that so far only a few tools (e.g., OCR, morph analyzer, morph generator, sandhi splitter, transliterator, a part-of-speech tagger) are developed for a few Indian languages. But we are not sure about the applicational potential of these tools as most of them are not easily accessible to general people. Therefore, we like to argue that we need to devise many corpus processing tools and techniques for most of the Indian language corpora keeping in mind the nature of Indian languages as well as the requirement of the language users. Here, however, we like to discuss some of the corpus processing techniques that we have used on the Bangla text corpus. And we hope that these techniques can be effectively used for other Indian languages with modifications as and when required.

11.4.1 Part-of-Speech Tagging

It is a type of text annotation, which involves attachment of special codes relating to part-of-speech of words used in a corpus in order to indicate their specific lexico-syntactic features and functions they exert in a piece of text. This process is therefore also known as 'grammatical tagging.' The code of the part-of-speech that is assigned to a word in a sentence is known as a 'tag.' When this operation is carried out on a piece of text in a corpus (manually or automatically), it follows a scheme for tagging a part-of-speech to each word used in a sentence. Normally, it is done at the following four stages:

(a) Pre-editing stage,
(b) POS determination stage,
(c) Tag assignment stage, and
(d) Post-editing stage.

At the pre-editing stage, a corpus is converted into a suitable format to assign a part-of-speech to each word or a word combination in a sentence. At the POS determination stage, form and function of a word are analyzed to determine its contextualized part-of-speech. At the tag assignment stage, each word is assigned that particular part-of-speech to which it should belong due to its role in the text.

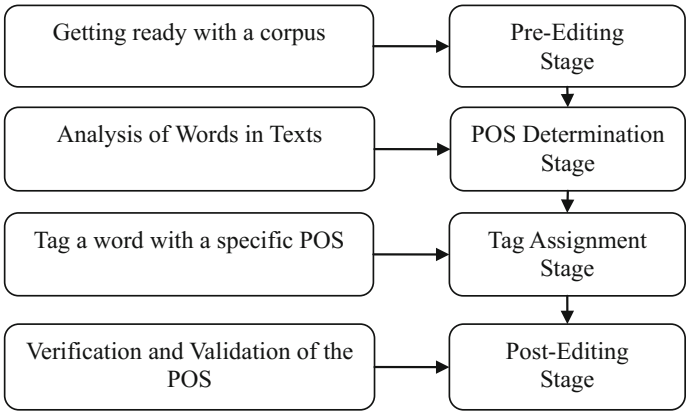

Fig. 11.1 Different stages of POS tagging in a corpus

At this stage, due to orthography, lexical role, and ambiguity, a word may appear to have more than one tag (e.g., cook_{NN} vs. cook_{FV}). The conflict of ambiguity has to be dissolved at this stage. It may be carried to the next stage if we find that more than one reading is possible due to the variation of part-of-speech of a word (e.g., *Time flies like an arrow*). The final stage is the post-editing stage, where each word, after being assigned with a particular tag, is checked by a man or a machine for verification and validation (Fig. 11.1).

At the initial stage of POS tagging, it is better to go with the primary list of part-of-speech available in a language without considering the finer subclassification of the major parts-of-speech as it may create much debate and confusion among the members engaged in POS tagging. Finer sophistication may be introduced with multilayered POS tagging at a hierarchical level once a corpus is done with initial POS tagging and made available for analysis and verification. Therefore, at the first trial of POS tagging, it is rational to adopt 'one-tier tag assignment' scheme, which can be done following the regular parts-of-speech of a language as shown in the list below (Table 11.2).

We have given below an example of POS-tagged text obtained from the Bangla text corpus (Fig. 11.2). Adopting the 'one-tier tag assignment' scheme, the text has been tagged following the primary tag set presented above (Table 11.2).

When we check through the last stage (i.e., post-editing stage) of the Bangla text corpus, we find that there are a large number of compound nouns, compound adjectives, adverbs, and reduplicated forms where the constituting words are physically detached from each other either by a space or by a hyphen (e.g., *bhul baśata* 'by mistake,' *sarkār niýukta* 'government-appointed,' *bhāsā prem* 'language love,' *mane mane* 'in mind,' *bāstab sammata* 'realistic'). The situation becomes far more complicated when we find that in case of a compound word made with two words (W_1 and W_2), the second word (i.e., W_2) carries an inflection marker while the first word (i.e., W_1) is used without an inflection (e.g., *basat bāṭir* 'of residential house,'

Table 11.2 One-tier tag assignment with regular parts-of-speech

No.	Part-of-speech	Tag	Example	Gloss
1	Noun	\NN\	bālak\NN\	Boy
2	Pronoun	\PN\	āmi\PN\	I
3	Adjective	\ADJ\	bhāla\ADJ\	Good
4	Adverb	\ADV\	kadācit\ADV\	Rarely
5	Finite verb	\FV\	karlām\FV\	Did
6	Non-finite verb	\NFV\	kariyā\NFV\	Doing
7	Postposition	\PP\	kāche\PP\	Near
8	Indeclinable	\IND\	kintu\IND\	But
9	Punctuation	\PNC\	?\PNC\	Interrogative

Bāṅglādeś\NN\ o\IND\ Bāṅgālī\NN\ jāti\NN\ samparke\PP\ paṇḍitgaṇ\NN\ e\ADJ\ parýanta\PP\ ýā\PN\ kichu\ADJ\ ālocanā\NN\ karechen\FV\ tār\PN\ theke\PP\ āmrā\PN\ ei\ADJ\ satye\NN\ upanīta\ADJ\ hayechi\FV\ ýe\IND\ bartamāne\ADV\ ýārā\PN\ mātṛbhāṣā\NN\ hisebe\PP\ Bāṅglā\NN\ bhāṣā\NN\ byabahār\NN\ karen\FV\ ebaṃ\IND\ Bāṅgālī\NN\ hisebe\PP\ paricita\ADJ\ tārā\PN\ bibhinna\ADJ\ nṛgoṣṭhīr\NN\ mānuṣ\NN\ niye\NFV\ gaṭhita\ADJ\.

Fig. 11.2 POS-tagged text obtained from a Bangla text corpus

mājh dariyāy 'in the mid stream,' *khabar kāgajer* 'of newspaper,' *baṛa lokder* 'of rich people,' *paścim bāṅglāte* 'in West Bengal'). In such cases, it is really tough to decide the actual part-of-speech of the words used to constitute the compounds.

If we run a lemmatization (discussed in Sect. 11.4.4) process before the POS tagging process is run on a corpus, all the two-word or multiword units will be decomposed into separate lexical units due to which the W_1 and the W_2 will be listed as separate lexical items. This will be a distortion of actual language data as well as non-reliable representation of lexical information in a language. This implies that unless the multiword units are tagged beforehand along with single-word units, there is a high chance of having serious mistakes in the process of lemmatization.

If we look into the POS-tagged sample presented above (Fig. 11.2), we can find some sets of important information about the text which are presented below:

(1) Each word in the text is identified with a specific POS tag.
(2) Words are tagged in accordance with their grammatical role in the sentence. If this is not done beforehand, we may fail to understand the actual syntactic and semantic role of a word in the text.
(3) If words are not previously tagged at the part-of-speech level, it will not be possible to identify their correct POS tag as well as their contextual meaning. This can have an adverse effect on the process of lemmatization.

(4) Some words, due to orthographic forms, may be identified as words of different POS if these are detached from the context of their occurrence. For instance, *samparke* can be tagged as \PP\ or \NN\, *ýā* can be tagged as \PN\ or \FV\, *tār* can be tagged as \PN\ or \NN\, *theke* can be tagged as \PP\ or \NFV\, *ýe* can be tagged as \IND\ or \PN\, *bartamāne* can be tagged as \ADV\ or \NN\, *hisebe* can be tagged as \PP\ or \NN\.

(5) In the POS-tagged text, there is an example of a compound noun where the formative words are written as two separate lexical items (e.g., Bāāṅglā$_{\backslash NN\backslash}$ bhāṣā$_{\backslash NN\backslash [NN_CMP]}$). If this compound word is not identified before and tagged properly as a single-word unit, there is a chance that this will be decomposed into two separate words at the time of lemmatization. And as a result of this, the actual lexical information of the compound word will be lost and the lemmatization process will yield wrong outputs.

These issues lead us to argue for implementing POS tagging on a corpus before the words are put to the process of lemmatization.

11.4.2 Concordance

The concordance is a process of indexing words used in a piece of text (see Chap. 5 of this volume). It is indispensable in the lexical analysis as it gives better scope to access possible patterns of use of words within a piece of text. It enables us to display the total list of occurrence of a lexical item—each occurrence in its own contextual environment (Dash 2007a, b, c, d, e). In concordance, words are indexed with close reference to the place of their occurrence in a piece of text to show their possible range of usage varieties in the text. Also, it can help us understand the distributional and semantic patterns of a word collected from a corpus in a desired manner for subsequent analysis and observation. The introduction of the computer has made concordance an easy process to compile and arrange words in the manner we desire. Due to flexibility in the technique, determination of contextual frame of words may vary depending on various criteria, e.g., fixed number words on either side of the target word (TW), finding the sentence boundaries of the TW.

The application of concordance on a corpus yields varieties of information, which are not available via intuition. Due to excellent advantage, it is used on corpus to search single and multiword units as well as scientific and technical terms along with details of contexts of their occurrences. With the help of concordance, it is not difficult to examine all the varieties of occurrence of different scientific and technical terms in a corpus. In the list below (Table 11.3), we have cited a sample concordance list of *software* to show how the term is used and how it varies in senses due to different contexts.

A well-designed concordance list can help terminology database developers to access and understand scientific and technical terms in their syntagmatic and paradigmatic frames. With options open for left- and right-hand sorting, it becomes useful

Table 11.3 Concordance of *software* taken from an English text corpus

the application of	software	such as a word processor
Firmware is a	software	that is programmed to
methods to test that a	software	is a fit product before it is
utilities and application	software	that serve in combination
for direct application	software	or subsets thereof we need
in computer technology	software	are often regarded as one
these types of	software	include web pages and all
at the lowest level	software	consists of machine reading
in computer science	software	engineering software is all
basis for most modern	software	was first proposed by
on generally used	software	systems on the desktop
computer system divide	software	system into three classes
the purpose of systems	software	is to unburden applications
the programming	software	usually provides tools and
there are three layers of	software	performing variety of tasks
usually a platform	software	often comes bundled with
to change the platform	software	the operational modalities

for terminologists to investigate if a technical or a scientific term is polysemous in nature with a wide range of sense variations. Thus, it gives direct access to the terms so that the terminologists can build profiles of meanings and uses of the terms. In essence, a printout of concordance lists offers a unique resource for perceiving similarities and differences of the scientific and technical terms in linguistic testing, analysis, and documentation.

11.4.3 Collocation

Collocation is a well-known linguistic phenomenon often discussed with evidence carefully selected from many languages. It is defined as 'occurrence of two or more words within a short space of each other in a text' (Sinlciar 1991: 170). The technique for identifying lexical collocation in a piece of text is important for evaluating the value of consecutive occurrence of any two words in a piece of text. In return, it projects into the functional nature of the lexical items used in a language as well as on the 'interlocking patterns of the lexis' in a text (Williams 1998). While studying the use of scientific and technical terms in a language, we are interested to know to what extent the actual patterns of use of these terms differ from the patterns that have been expected to form (Barnbrook 1998: 87). This query is related to the argument that claims that our mental lexicon is made up of not only with single-word units but

also with a larger number of multiword units, both fixed and variable. Therefore, a reference to the collocational patterns of scientific and technical terms can give us better insights to comprehend their forms, associations, functions, and implications.

A collocation program, when it runs on a corpus, produces various kinds of information about the nature of collocation of words and terms used in a text. A systematic analysis of collocation can help us understand the position and function of the scientific and technical terms that frequently take part in collocation in a language. Thus, a list of collocation of scientific and technical terms obtained from a corpus carries vital information about their patterns of association which we need to understand to analyze their nature of sense denotation.

A list of collocation may include information about the frequency of use of scientific and technical terms in collocation as well as specific statistical counts for calculating their frequency patterns in collocation (Dash 2008). In fact, without adequate reference to their frequency of use in collocation, we cannot understand the finer aspects relating to their patterns of distribution, the possible range of sense variation, and distinctions among the senses they denote when they are used in different contexts.

The analysis of collocation of terms shows that by referring to contexts we can empirically determine which pair of scientific and technical terms maintains substantial collocation relationship between them. The most suitable formula used is the mutual information scale (MIS) that helps to compare the probability of any two scientific or technical terms (STT_1 and STT_2) occurring together as an event with their probability of occurrence as a result of chance. For each pair of terms, we take a statistical score from a corpus to conclude that where there is a higher score, the greater is the possibility of their collocation. Thus, a reference to MIS becomes necessary in the evaluation of the patterns of occurrence of collocation of scientific and technical terms used in a language.

The functional relevance of collocation in the areas of terminology database generation, research, and application is many, some of which are listed below:

(1) Information of collocation helps to extract multiword scientific and technical terms from a corpus to compile separate databases of scientific and technical terms as well as databases of translational equivalents.
(2) It helps in analyzing the patterns of collocation of scientific and technical terms as well as to design materials for teaching technical texts.
(3) It helps to group all the multiword scientific and technical terms in a separate list to identify the range of their sense variation as well as to know how they generate new senses by collocating with other words.
(4) Understanding patterns of collocation of scientific and technical terms helps us understand and identify their differences in use in a piece of text.
(5) It helps to understand the nature and pattern of semantic linkage between the two synonymous scientific and technical terms.
(6) Patterns of collocation of scientific and technical terms obtained from a corpus show that they may have vital differences in lexical associations resulted from the difference in distribution across discourse types.

In essence, the analysis of examples of collocation of scientific and technical terms can show that they are rarely equivalent in sense and function when considered in terms of their distribution in a piece of text. Thus, information regarding delicate differences of collocation of scientific and technical terms becomes important inputs for the people engaged in terminology database generation, dictionary and TermBank compilation, machine-aided translation, speech and language processing, dictionary compilation, and language teaching.

11.4.4 Lemmatization

In traditional linguistics, the term 'lemma' refers to the basic form of words disregarding their grammatical changes such as tense and plurality (Biber et al. 1998: 29). The process of lemmatization is related to the identification of parts-of-speech of words used in a piece of text and reducing the words to their respective lexemes—the headwords that we look for in a dictionary (Dash 2007a, b, c, d, e). For many tasks relating to corpus processing such as frequency counts, alphabetical sorting, collocation analysis, lemmatization is an indispensable technique in which we can group together different forms of an inflected word so that we can collectively display them under one head (Barnbrook 1998: 50). In the area of vocabulary study and dictionary compilation, it allows us to produce frequency and distribution information for the lemmas (Sánchez and Gomez 1997). Given below is an example of lemmatization taken from a Bangla text corpus (Table 11.4). It shows how the inflected scientific and technical terms can be easily lemmatized to generate a database of the same.

Input Text:

svādhīnatā lābher par theke gata calliś bachare kendrīya sarkār katakguli bhrāntimūlak nīti anusaraṇ kare esechen.

The process of lemmatization allows us to extract and examine all the variants of a particular lemma and to produce detailed frequency and distribution information for them in a separate list. In case of compiling a database of scientific and technical terms, it helps us capture the possible number of inflected and affixed forms generated from a particular scientific or technical term. It also guides us understand which terms are inflected, how many times these are inflected, and in which manners these are inflected.

11.4.5 Frequency Sorting

The frequency of use of scientific and technical terms in a corpus is of great value in the collection of scientific and technical terms as well as in the development of scientific and technical terminology database. In the process of frequency sorting, scientific and technical terms are arranged in accordance with their frequency of

Table 11.4 Lemmatization of words in a Bangla text corpus

Surface form	Part-of-speech	Base/root	Suffix
svādhīnatā	NN	svādhīnatā	–
lābher	NN	lābh	-er
par	NN	par	–
theke	PP	theke	–
gata	ADJ	gata	–
calliś	ADJ	calliś	–
bachare	NN	bachar	-e
kendrīya	ADJ	kendrīya	–
sarkār	NN	sarkār	–
katakguli	ADJ	katak	-guli
bhrāntimūlak	ADJ	bhrāntimūlak	–
nīti	NN	nīti	–
anusara ṇ	NN	anusaraṇ	–
kare	NFV	kar	-e
esechen	VB	es (<ās)	–

occurrence in a corpus to identify which scientific and technical terms are more frequent and which are least frequent in use. The list can be made in ascending as well as in descending order based on the needs of the terminologists.

Generally, a small-sized and less representative corpus can provide too small a list of scientific and technical terms to be interesting and useful. But a large-sized, multitextual, and widely representative corpus of billions of words can be highly relevant in producing frequency lists that might be useful for gathering terminological information in a language, as the listed terms become comparable to a large population for statistical authenticity. Since the most frequently used terms show varieties in their occurrence in texts, many marked changes in the patterns of their distribution become significant in linguistic analysis and generalization. For instance, while highly frequent terms are attested even in a small corpus, less frequent terms will not occur unless a corpus is large enough with a large number of samples obtained from various text types.

Since a frequency list provides necessary clues to know which scientific and technical terms occur in which frequency in a language, by examining a frequency list, we can collect the most frequently used terms to compile graded terminology databases for language teaching. Also, information on the frequency of use of scientific and technical terms becomes useful inputs in dictionary compilation, terminology database generation, and term-based text compilation. In case of developing dictionaries of limited terms, frequency information leads us to decide which terms will be included, since the most frequent terms normally get priority over the rarely used ones. Thus, the frequency of use of scientific and technical terms leads us to

enlist and accept the most frequent forms so that we can elaborate them to make them popular among the common users.

We can also use frequency information to measure the meaning variation of scientific and technical terms. Since the most common meaning of the terms occurs more often than their least common meaning, this becomes relevant in identifying the senses of scientific and technical terms as well as in selecting the most frequent senses for reference and use in dictionaries and language teaching resources (Wills 1990: 142). In essence, frequency information of senses establishes its importance in scientific and technical terms analysis, description, sense disambiguation, and documentation.

Since the majority of Indian language corpora are not yet used for frequency-based analysis for collecting scientific and technical terms, we think we should give some attention toward this area. However, when we do this, we should keep in mind the problems relating to identification and analysis of scientific and technical terms used in the Indian language corpora. Else, we shall make false observation and wrong deduction about these terms and their linguistic information.

11.4.6 Type–Token Analysis

The last and the most indispensable stage of corpus processing is the type–token analysis of scientific and technical terms occurring in a corpus. It is made up of two basic steps:

(a) Alphabetical sorting of the terms and
(b) Removal of multiple tokens after storing a type of a term.

Once the scientific and technical terms are obtained from a corpus, they are passed through the stage of alphabetical sorting so that the terms are arranged in alphabetical order. The alphabetical sorting is a list of the terms, which are arranged in alphabetical order with a tag denting their frequency of use in the corpus. Since it is used for simple general reference purposes, it usually plays a secondary role in the context of checking the frequency of use of the particular term in the text. However, it can be useful as an object of independent study as it helps us in the formulation of hypotheses as well as for checking the assumptions that have been made before (Kjellmer 1984: 9). At the time of alphabetical sorting, scientific and technical terms are displayed in a vertical form for general reference purposes and the list is formed in such a way that each term is put in a separate line for better comprehension.

The next stage is related to the removal of multiple tokens from the list of terms. Since an alphabetically sorted list contains multiple entries of the same term (i.e., tokens), we need to remove the identical terms from the list after storing one of the variants (i.e., types) as a representative of the tokens. For instance, an alphabetically sorted list may contain ± 100 tokens of a single term (say, *printer*), either in its inflected form (e.g., *printers*) or non-inflected one (e.g., *printer*). The primary task is, therefore, to preserve only one of the tokens as a 'type' and remove the other

forms from the list. Other forms are removed because they are identical replicas of the type selected for the list. By this process, a large list of tokens may be reduced to a small and manageable set of types, which may be preserved in the final database of scientific and technical terms for a language.

11.5 Scientific and Technical Term Database

While developing a database of scientific and technical terms, we have to distinguish between two types of terms, based on the nature of accumulation:

(1) **Ad Hoc Database**: This will contain only the most frequently used scientific and technical terms with a goal for using them in translation as well as in various other works carried out by the professionals. The translators, while translating scientific and technical texts, may ask for specific terms or group of terms to solve particular translation problems.
(2) **Complete Database**: It will contain almost all scientific and technical terms found to be used in a particular subject area or in all domains of human knowledge. It should be exhaustive and complete as far as the availability of the terms of a subject domain is concerned.

We can postulate three methods in the manner accumulation of scientific and technical terms:

(1) **Selection Method**: In this method, we can select limited scientific and technical terms from books and journals of different fields and disciplines, covering physical and natural sciences, geosciences, social sciences, engineering, technology, medicine, commerce and business, art and humanities, administration, law and legal documents, etc. We can select them manually to compile a small database. In this case, however, we have to go through different types of text to pick up the terms we think appropriate to be included in the database.
(2) **Collection Method**: In this method, we can collect scientific and technical terms from earlier databases already available in a language, such as dictionaries, word books, books of terms, and word books of specific fields and disciplines. Here, although our task is not exhaustive, yet it will ask for a careful collection of the terms suitable for the database.
(3) **Holistic Method**: This is the most useful method of scientific and technical term database compilation. Here, we first collect terms from the other two methods stated above. Next, we process corpora of different disciplines, domains, and fields in multiple ways (as discussed in the previous section) to collect scientific and technical terms with all necessary information required to be furnished within a terminology database.

The primary activities relating to the extraction of scientific and technical terms from a digital corpus is presented in the following diagram (Fig. 11.3).

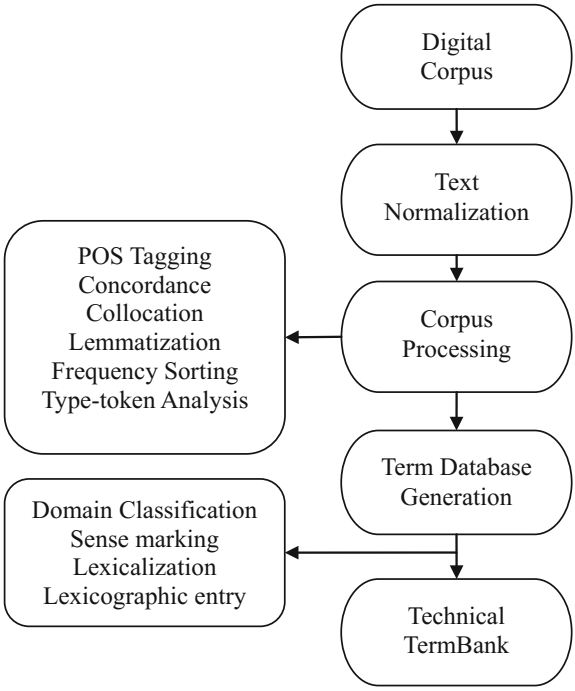

Fig. 11.3 Technical TermBank generation from a corpus

Finally, after compilation of an exhaustive list of scientific and technical terms from several domain-based corpora following the methods and strategies discussed in this chapter, it is necessary to develop a TermBank in digital form for all purpose usages. The most important questions are related to the storage of the terms and the load of information to be attached to the terms. Since the final TermBank is meant to be developed in digital form, there is hardly any issue with a paucity of storage space. We can store all scientific and technical terms as we find them from different domain-specific text corpora of a language. Also, we can store as much as we find suitable for a TermBank. Now, the question is related to the amount and types of lexicographical data and information to be furnished with each term in the TermBank. The following diagram (Fig. 11.4) can give us a clear idea about the amount and type of information to be tagged with each term stored in the TermBank.

A TermBank should, in principle, possess the above sets of information. A text corpus or a combination of several text corpora will provide the necessary information required for each of the terms, except the information relating to origin or etymology. This may be availed from our structured linguistic knowledge sources like dictionaries and grammars.

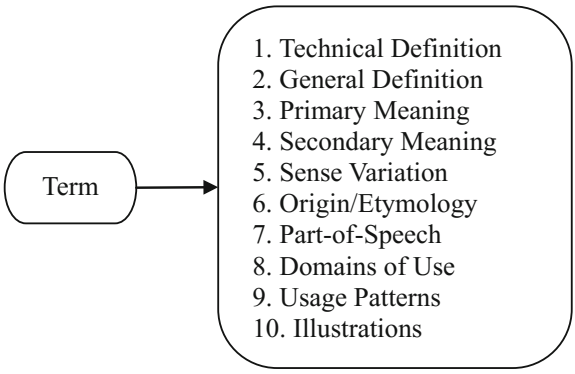

Fig. 11.4 Information to be used with each term in a TermBank

11.6 Conclusion

We can visualize four types of users of scientific and technical term databases: language specialists, content specialists, language technologists, and general people. Among the language specialists, the dictionary makers will require detailed information about the general and specific uses of the scientific and technical terms to develop term dictionaries, subject-specific dictionaries, and reference dictionaries. The terminologists and technical writers will need data and information to standardize the technical terminology database as well as to increase the existing terminology database of a language. They use these terms to investigate linguistic phenomena of diverse kinds and verify evidence of their own or others. The language teachers, as well as a learner, will refer to this database at the time of teaching, learning, writing course books, and similar works.

Among the content specialists, historians will require this database to carry diachronic studies on a language through elaborate analysis of scientific and technical terms used in earlier texts. They also use this database to discover implicit marks of time recorded in the terms of obscured documents. The literary critics will use the terms in their research into stylometrics, as statistical information about the use of scientific and technical terms can play a crucial role in identifying authors of dubious texts. They can also use the terms stamped with statistical information to identify different types of text based on the density of use of the terms.

Among the language technologists, people engaged in information retrieval can use term databases to devise mechanisms for extracting information from the large body of texts to build a lexical knowledge base, find information of terms for indexing, and to summarize the important content of texts. Also, they can use the scientific and technical terminology database to test the presence or absence of regularities of use of the terms in a text. The people engaged in machine translation can utilize the term databases as necessary inputs to develop bilingual and multilingual translational equivalents that may be used in manual and machine translation.

Finally, general people can use the terminology database in language description, language study, text composition, language cognition, language therapy, and publication.

References

Barnbrook, G. 1998. *Language and Computers*. Edinburgh: Edinburgh University Press.
Barnhart, C. 1978. American Lexicography. *American Speech*, 83–140.
Biber, D., S. Conrad, and R. Reppen. 1998. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.
Dash, N.S. 2005. Role of Context in Word Sense Disambiguation. *Indian Linguistics* 66 (1–4): 159–175.
Dash, N.S. 2007a. Generating Electronic Lexical Resources from Text Corpora in Bengali. *Indian Linguistics* 68 (3–4): 361–371.
Dash, N.S. 2007b. Indian Scenario in Language Corpus Generation. In *Rainbow of Linguistics*, ed. by N.S. Dash, P. Dasgupta, and P. Sarkar, vol. I, 129–162. Kolkata: T. Media Publication.
Dash, N.S. 2007c. Some Techniques Used for Processing Bengali Corpus to Meet New Demands of Linguistics and Language Technology. *SKASE Journal of Theoretical Linguistics* 4 (2): 12–31.
Dash, N.S. 2007d. Toward Lemmatization of Bengali Words for Building Language Technology Resources. *South Asian Language Review* 17 (2): 1–15.
Dash, N.S. 2007e. *Language Corpora and Applied Linguistics*. Kolkata: Sahitya Samsad.
Dash, N.S. 2008. Techniques of Text Corpus Processing. In *Readings in Quantitative Linguistics*, ed. P. Mohanty and R. Köhler, 81–115. Indian Institute of Language Studies: New Delhi.
Dash, N.S., and P. Basu. 2012. Developing Scientific and Technical Terminology Database from Electronic Language Corpora. *Language Forum* 38 (1): 5–21.
Kjellmer, G. 1984. Why 'Great: Greatly' But Not 'Big: Bigly? *Studia Linguistica* 38: 1–19.
Sánchez, J.A., and P.C. Gomez. 1997. Predictability of Word Forms (Types) and Lemmas in Linguistic Corpora, a Case Study Based Analysis of the CUMBRE Corpus: An 8-Million-Word Corpus of Contemporary Spanish. *International Journal of Corpus Linguistics* 2 (2): 259–280.
Sinclair, J.M. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
Williams, G.C. 1998. Collocational Networks: Interlocking Patterns of Lexis in a Corpus of Plant Biology Research Articles. *International Journal of Corpus Linguistics* 3 (1): 151–172.
Wills, J.D. 1990. *The Lexical Syllabus*. London: Collins.

Web Links

https://legal-dictionary.thefreedictionary.com/Scientific+terms.
http://www.whatislife.com/glossary.htm.
https://www.edge.org/annual-question/what-scientific-term.
https://en.wiktionary.org/wiki/technical_term.
http://www.yourdictionary.com/technical-term.
https://www.definitions.net/definition/technical%20term.
https://legal-dictionary.thefreedictionary.com/technical+term.
http://web.mit.edu/course/21/21.guide/techterm.htm.
https://corpus-analysis.com/.
http://www.laurenceanthony.net/software/antconc/.
http://tdil-dc.in/index.php?lang=en.

Chapter 12
Corpus and Machine Translation

Abstract History shows that a machine translation (MT) system with the support of a few linguistic rules is not realistic. A few rules are not sufficient for capturing the wide variety a natural language exhibits in its diverse use. This leads us to argue for a corpus-based machine translation (CBMT) system that desires to rely on a large amount of linguistic data, information, examples, and rules retrieved from corpora. The first benefit of a CBMT system is the development of algorithms for alignment of bilingual text corpus (BTC)—an essential part of an MT system. A BTC generates a new kind of translation support resource that helps in learning through trial, verification, and validation. A CBMT system begins with analysis of translations produced by human to understand and define the internal structures of BTC, completely or partially, to design strategies for machine learning. Analysis of BTC lends heavily to develop aids to translation as we do not expect an MT system to 'produce' exact translation but to 'understand' how translations are actually produced with linguistic and extralinguistic information. The use of BTC in CBMT is justified on the ground that data and information acquired from BTC are richer than monolingual corpus with regard to information of contextual equivalence between the languages. Thus, a CBMT system earns a unique status by a combination of features of the example-based machine translation (EBMT) and statistics-based machine translation (SBMT) keeping a mutual interface between the two.

Keywords Bilingual translation corpus · Text alignment · Text normalization Translational parallelism · Translational equivalence · Lexical matching Grammatical mapping · Syntactic rules

12.1 Introduction

By a simple definition, machine translation (MT) is a technique that takes input in the form of full sentences from a source language (SL) and generates a corresponding full sentence in a target language (TL). The MT technology has improved a lot with good availability of data, information, and technology. We consider it now as a key

N. S. Dash and L. Ramamoorthy, *Utility and Application of Language Corpora*,
https://doi.org/10.1007/978-981-13-1801-6_12

technology that can have a lasting impact on the global commercial market of cross-lingual information, interlingual communication, and information exchange. MT is now a cross-disciplinary field with direct impact on e-commerce, localization, and knowledge-based society (Winograd 1983).

Conceptually, the corpus-based machine translation (CBMT) system is based on a range of issues evolved from empirical analysis of BTC. The act of analysis involves both linguistic (morphological, semantic, and functional interpretation of words, terms, phrases, sentences, paragraphs, etc.) and extralinguistic analyses of data and information present in a BTC. It employs various statistical techniques to generate probability statistics from BTC to identify nearest possible translation equivalents from the two languages (Altenberg and Aijmer 2000). It stands on the assumption that there are no pre-established solutions to an MT, but possible solutions may be found from analysis of manually translated texts stored in a BTC. It assimilates and applies a large amount of competence of human translators encoded in the linguistic equivalents found in BTC. Success achieved from restricted domains leads us to hope that linguistic and extralinguistic information found from BTC are essential for achieving success in the general domain.

At present, it is too early to make any prediction that a CBMT system will succeed in all domains of human knowledge. However, we may argue that an MT system developed with data and information taken from a well-representative BTC can be more robust and useful both in restricted and general domains. In essence, a CBMT system which is based on data, information, and examples gathered from analysis of BTC can make the dream of MT a reality in years to come.

During the last seven decades, history has taught us that designing an MT system with a set of rules is not realistic. A set of rules is not enough to encompass a wide variety of a language exhibited in diverse domains. This leads to the birth of corpus-based machine translation (CBMT) system that combines both the example-based machine translation (Furuse and Lida 1992; Jones 1992; McLean 1992; Somers 1999; Dietzel 2009) and the statistics-based machine translation (Brown et al. 1990; Brown et al. 1993; Koehn 2010) to reach to a stage still elusive in MT. Let us see how a CBMT system wants to reach that destination.

The chapter is organized as follows. In Sect. 12.2, we focus on the basic issues relating to CBMT; in Sect. 12.3, we discuss the process of creating a translation corpus in the languages involved in MT; in Sect. 12.4, we focus on the process of text alignment in BTC; in Sect. 12.5, we highlight the basic linguistic tasks on a BTC; in Sect. 12.6, we address the process of analysis of BTC; in Sect. 12.7, we argue for building a bilingual dictionary; in Sect. 12.8, we discuss extraction of translational equivalents from BTC; in Sect. 12.9, we propose for generating a TermBank; in Sect. 12.10, we discuss the process of lexical selection from TL; in Sect. 12.12, we highlight the problems of grammatical mapping; in Sect. 12.13, we refer to other issues involved in CBMT; in Sect. 12.14, we present a system module to be adopted for a CBMT system.

12.2 Issues of a CBMT System

The idea of using BTC in CBMT is not a new thing. It dates back to the early days of MT, but it was not used in practice until 1984 (Kay and Röscheisen 1993). A CBMT system is based on information and data acquired from analysis of BTC because a BTC is more enriched with information about structural similarities of the languages than a monolingual corpus. Also, a BTC provides information of situational equivalence on the possibilities of a language system when it comes in contact with a different language system. It is based on a range of information obtained from empirical analysis of BTC (Baker 1996). Analysis can be of different types: morphological, syntactic, semantic, lexical, phrasal, idiomatic, figurative, pragmatic, discoursal, semiotic, and extralinguistic. In the analysis, it employs various statistical methods and techniques to generate probability measurements from BTC to identify translational equivalents for the languages.

A CBMT system stands on the assumption that there are no pre-established solutions to translation, but possible solutions may be found in those texts, which are translated by human translators. In other words, a large portion of competence of human translators is encoded in language equivalence found in translated texts (Su and Chang 1992). Success achieved in this method in restricted domains leads us to argue that both linguistic and extralinguistic information are essential for success in the general text (Teubert 2002).

In recent times, a CBMT system has made considerable advancement through extensive analysis and research of BTC. BTCs are developed in many languages and are analyzed to gather important insights to design useful techniques. A BTC represents a large collection of real empirical texts collected through samples that reflect on the needs of end users. This factor becomes important for those end users who want to select a system that can translate texts to suffice their specific needs. It is clear that utilization of BTC is necessary since it supplies numerous linguistic and extralinguistic data and information to make a system robust. The primary issues relating to the development of a CBMT system are the followings:

(a) Generation of bilingual translation corpus,
(b) Alignment of bilingual translation corpus,
(c) Linguistic analysis of bilingual translation corpus,
(d) Extraction of translational equivalents,
(e) Generation of terminology data bank,
(f) Building a bilingual dictionary,
(g) Algorithm for lexical selection,
(h) Dissolving lexical ambiguity,
(i) Formation of grammatical mapping rules,
(j) Formation of lexical mapping rules,
(k) Selection of translational equivalents,
(l) Addition of pragmatic information,
(m) Addition of sentential information.

One of the first results received from a CBMT system is the development of algorithms for aligning sentences in a BTC. It is one of the major issues in MT as it constitutes in itself a suitable foundation for a new kind of translation support. A CBMT system begins with translations already produced by human translators to discover similarities in the internal structures of the languages. This analysis-oriented perspective lends to the development of translator's aids as a CBMT system, at its early stage, is not expected to 'produce' translation, but 'understand' them to become useful in subsequent stages.

12.3 Creation of Bilingual Translation Corpus

A BTC has two parts: original texts from the SL and their translation in the TL. A BTC usually keeps meaning and function of words and phrases constant in both the ST and TT to offer a scope for comparing meanings of the lexical stock of the two languages under identical condition (Koehn 2005). It helps to discover crosslinguistic variants, i.e., alternative forms of particular meanings and concepts, in the TL. Thus, a BTC provides useful resources for cross-lingual mapping and rule formulation for a translation system (Altenberg and Aijmer 2000: 17).

The creation of a BTC is, however, a challenging task (Dash and Arulmozi 2016). It requires constant careful monitoring from experts corpus linguists having good exposure in corpus generation and processing. A diagram is given below (Fig. 12.1) to show that a BTC can be developed between SL and TL in an organized manner following certain rules. After creation, it can be used as a comparable corpus (A :: C); as a bidirectional translation corpus (A :: B and C :: D); as a source for comparing original and translated texts in the same languages (A :: D and C :: B); and for comparing translated texts in both languages (A :: C and B :: D).

A BTC combines advantages of the comparable and parallel corpus. The texts in SL and TL match as far as possible in terms of genre, subject area, purpose,

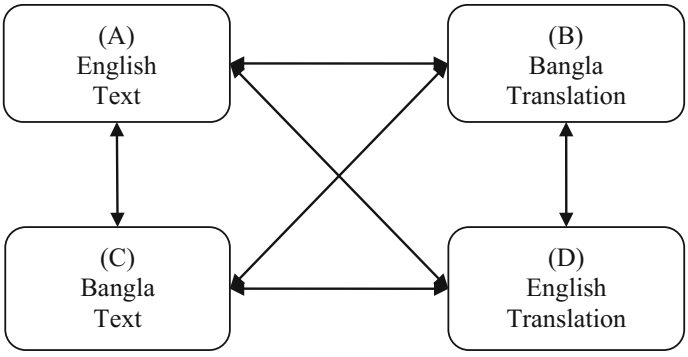

Fig. 12.1 BTC (Altenberg and Aijmer 2000: 17)

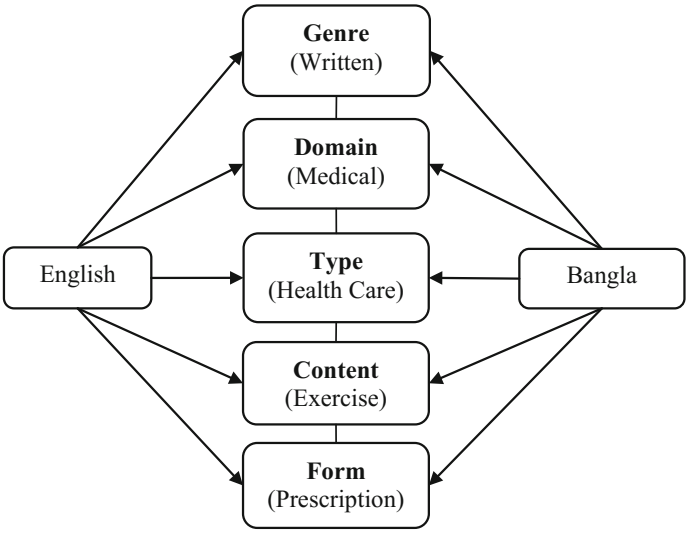

Fig. 12.2 Bilingual bidirectional parallel translation corpora

and register (Sanderson and Croft 2012). Samples of written texts are included in a BTC as bilingual translated speech corpus is not of much use in this task. Moreover, samples from spoken texts are not included since the present MT research targets written texts only. The included texts are expected to reflect on contemporary features and aspects of SL and TL, although older texts may have relevance in the translation of historical texts. Since a BTC is not limited to the texts of specific disciplines, it includes a wide range of texts from all domains and disciplines of language use. The composition and structure of a BTC are envisaged as follows keeping in mind the basic components of SL and TL texts (Fig. 12.2).

The text samples of SL and TL should maximally match. They should match in the genre (e.g., written), domain (e.g., politics), text type (e.g., news), content (e.g., election), and form (e.g., report). They should also match on purpose, type of user, subject matter, and register. Text samples in a BTC should contain fairly large and coherent extracts from beginning to end at a natural breaking point (e.g., paragraph, section, chapter). The diagram above shows that a BTC can also be used as a comparable and bidirectional corpus for various linguistic tasks (Fig. 12.2).

12.4 Alignment of Texts in BTC

Aligning text in a BTC means making each translation unit (TU) of SL correspond to an equivalent unit in TL. A TU may cover all short units like words and terms; medium units like phrases, idioms, and sentences; and large units like paragraphs

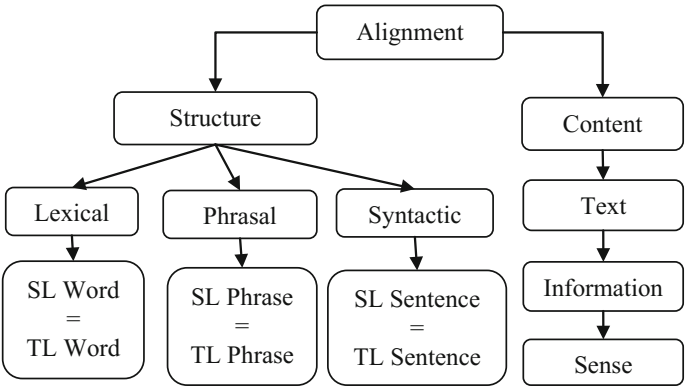

Fig. 12.3 Alignment of texts in a BTC

and chapters. At any level of text alignment (e.g., word, sentence, paragraph), a BTC is considered as a pair of texts with 'parallel units' (Castillo 2010). The selection of TU depends largely on the point of view considered for linguistic analysis and the type of corpus used. If a BTC asks for a high level of faithfulness to the original (as happens in case of legal and technical texts), it is necessary to align TUs at word, phrase, and sentence level. On the other hand, if a corpus is an adaptation rather than literal translation of SL text, alignment at the level of paragraphs and chapters will suffice. Thus, the alignment process is defined based on the type of corpus used in translation. Authenticity and linearity of human translation may guide one to align a BTC, although this is mostly true to technical texts (Chen and Chen 1995). A literary text corpus may require alignment at the sentence and above level if the types of translation equivalence observed in the corpus are previously formalized (Fig. 12.3).

At the time of text alignment in a BTC, the following three possibilities are to be kept in mind:

(a) High accuracy between typologically and genealogically close languages,
(b) Medium accuracy between typologically and genealogically near languages,
(c) Low accuracy between typologically and genealogically distant languages.

At the initial phase, a BTC may be used to trace correspondence, if not equivalence, where contents of the texts and their mutual relationships are put under consideration. Such 'Free Translation' may, however, create a serious problem in BTC processing due to missing sequences, changes in word order, modification in content, etc. Although these are common in normal translations, their frequencies vary based on the domain of a text. These may lead one to produce aligned corpus which is not a set of some equivalent sequences, but a corresponding text with mutual conceptual parallelism (Chen and Chen 1995). The goal is not to show structural equivalence between two languages, but pragmatically to find those units in TL that is semantically close to the units in SL (Dash 2005).

The starting point is a preliminary alignment of words with a bilingual dictionary. Rough alignment yields satisfactory results at the sentence level when it is supported by some statistical methods with the minimum formalization of major syntactic issues. The main advantage of this method is the use of translation memory (TM) for integration of data found in BTC. The task may be further simplified by using a reference corpus of a specialized field (e.g., *medical texts, legal texts, computer science texts, physics texts, chemistry texts*) from SL and TL. The message is thus 'machine-translated' by using a customized dictionary to create TM during a training phase. In essence, alignment at the sentence level is an important part of translation analysis as it desires to mark correspondences down to the level of sentences between SL and TL.

A week translation alignment scheme may serve the purpose, as this is one of the basic resources required at the initial stage of BTC analysis. One needs to develop a translation analyzer (TA), which will account for translational correspondences between morphemes, words, idioms, and phrases of a BTC. Another challenge in translation analysis is the use of statistical techniques for searching equivalent items in a BTC. The statistical searching algorithm may use semantic information to find equivalent units from a BTC, and once these items found, these may be verified and formalized by experts as valid inputs of a bilingual TM. This process may be used to automatize the training phase as well as to validate translation outputs. This is one of the basic criteria that can mark out differences between automatic translation and an MT system supported by a BTC.

A crucial limitation of a CBMT system is the paucity of parallel BTC (in the Indian context) due to which the developers of a CBMT system have been crippled for ages. It is understood that information obtained from a parallel BTC helps developers design an MT system as well as test reliability of the system. We, therefore, argue for compiling parallel BTC not only to meet the research requirements but also to evaluate efficiency and usefulness of a CBMT system. It is also understood that information obtained from a parallel BTC can provide necessary cues to identify the patterns about how SL and TL texts are structurally interlinked to each other and how information encoded in SL text is transferred to TL text (Baker 1993). Also, a parallel BTC helps to improve the standard of a CBMT system by gathering new insights into intricate linguistic relations between the paired languages.

12.5 Linguistic Tasks on a BTC

A BTC, after compilation and alignment, is put to linguistic analysis. It is an important phase of text formalization for retrieving translation equivalents. It has several steps:

(a) Morphological analysis: To identify the form and function of constituting morphemes of words,
(b) Lexical analysis: To identify lexical identity (i.e., *origin, part-of -speech, surface structure,* etc.) of lexical items,

(c) Syntactic analysis: To identify the form and function of syntagms (i.e., *phrases, clauses,* etc.) in respective texts,
(d) Morphosyntactic analysis: To understand the morphosyntactic properties (e.g., *inflection, affixes, case markers*) tagged to lexical items. Accurate and effective analysis of it enhances the quality and speed of processing,
(e) Semantic analysis: To identify the meaning of words, terms, idioms, phrases, etc., and ambiguities involved therein,
(f) Extralinguistic analysis: To identify and record topic, co-text, context, discourse, pragmatics, culture, and other information.

For effective linguistic analysis, it is sensible to use a superficial and descriptive morphosyntactic approach (i.e., *part-of-speech tagging, shallow parsing,* etc.). Also, simple statistical approaches may be used for probability measurement. A text analyzer may be supported by a standard grammar acquired from previously processed corpora as these corpora transcend its application in subsequent stages. In an ideal situation, POS tagging may be performed in a supervised method by comparing text samples included in BTC following 'probabilistic' procedures. Although in this way, for example, some adjectives (ADJ) may be translated as nouns (NN) or vice versa (e.g., *good, bad, old, new, cold, red, dark*), the defined categories in regular grammars and dictionaries may help to resolve grammatical ambiguities.

At this stage, traditional parts-of-speech have good referential impacts on the quality of POS tagging, since a translation system with fewer parts-of-speech gives a better result than a system with elaborate parts-of-speech. Although here the main objective is to create some active translation memories, other utilities (e.g., *bilingual lexical database, bilingual paired sentence database, bilingual terminology databases, machine learning, and computer-assisted language teaching*) are also possible with this resource. A BTC can also be used as a resource for developing digital dictionaries and writing bilingual comparative grammars.

12.6 Analysis of a BTC

Within the domain of CBMT research, the central point of debate is the question about the levels of complexity involved in BTC analysis. The general assumption is that unless a large number of linguistic phenomena occurring in a BTC are analyzed and overtly represented, a good-quality CBMT system is not possible to design. We argue that problems like *lexical ambiguity* and *constituent mapping* may be solved with the knowledge obtained from a BTC and stored in lexicon and grammar of SL and TL. This, however, requires the application of a rigorous analysis scheme on a BTC to make explicit some or all the translation correspondences that link up the segments of SL texts with those of TL texts.

The BTC analysis technique may be used to structure the preexisting translated texts in such a manner that they are reusable for generating new translations. The BTC analysis technique may be used to draft the translations as well as to detect

various translation errors occurring in a BTC. Once translation correspondences are reconstructed between ST and TL texts, one can verify if correspondences have constraints of any kind. For instance, we can claim a translation complete if larger chunks (e.g., *pages, sections, paragraphs, sentences*) of SL text are properly translated in TL texts.

The BTC analysis technique may also be used to verify if a translation is free from *interference error* which is caused by 'deceptive cognates' or 'false friends.' For instance, the Bangla word *sandes* and the Hindi word *sandes* may appear same, but actually they are two different words with two different meanings. Similarly, Bangla *bau* and Odia *bau*, Bangla *sūcanā* and Hindi *sūcnā*, Bangla *khun* and Hindi *khun*, Bangla *pataṅg* and Hindi *pataṅg*, etc., may appear as 'false friends' in a Hindi-Bangla BTC. Such forms should not be accepted as appropriate cognates for mutual translation, although they may appear nearly identical in orthographic representation in the two languages.

12.7 Building a Bilingual Dictionary

An essential component of a CBMT system is a bilingual dictionary, the lack of which has been a great stumbling block in CBMT research for the Indian languages. Traditional dictionaries cannot make up this deficiency, as they lack in information about the types of lexical subcategorization, patterns of lexical distribution, nature of lexical selection restriction, domains of use of lexical items, etc. By using statistical methods, it is possible to extract lexical subcategorization information from a POS-tagged BTC and store them in a bilingual dictionary (Vandeghinste 2007). In that context when a POS-tagged BTC is not available, a bilingual dictionary produced from untagged BTC may be used for this task (Dash 2016).

The development of a bilingual dictionary is best possible within those languages which are typologically or genealogically related to each other (e.g., *Bangla: Odia*, *Bangla: Assamese, Hindi*: *Urdu*). Most of these languages usually share many properties (both linguistic and non-linguistic) which are common to both the languages. This kind of similarity is hardly found between the non-related languages. Also, there is a stock of regular vocabulary, which is similar to each other not only in orthographic structure and morphosyntactic form, but also in grammatical function, lexicographic meaning, contextual senses, and connotation. Given below is a list of words which show how many words are the same in form and sense between genealogically related languages like Bangla and Odia (Table 12.1).

For generating a bilingual dictionary, one can use statistical methods on a tagged BTC. The primary works for this are the followings:

(a) Retrieve large comparable syntactic blocks (e.g., *phrases, clauses*) from a BTC.
(b) Extract various subcategorized constituents (e.g., *subject, object, predicate*) from a tagged BTC.

Table 12.1 Similarity at word level between Bangla and Odia

Words	Bangla	Odia	Gloss
Pronouns	āmi	mu	I
	tumi	tume	You (regular)
	āpni	āpana	You (hon)
	tui	tu	You (non-hon)
Postpositions	kāche	pākhare	Near
	mājhe	majhire	Between
	nice	talare	under
	pāśe	pākhare	At
	upare	upare	Above
Indeclinable	ebang	madhya	And
	kintu	kintu	But
	ba	āu	Or

(c) Extract most frequently occurring adverbial clauses, adjectival phrases, idiomatic expressions, set phrases, etc.
(d) Identify lexical items considered as true translation equivalents due to similarities in form, meaning, and usage from BTC.
(e) Identify and extract identical proper names and pronouns from a BTC, and confirm their functional matching.

One should not expect hundred percent success in tracing similarity at morphological, lexical, syntactic, semantic, and extralinguistic levels in a BTC, even though the languages are closely related to each other. With all information extracted from a BTC, a *Core Grammar* is a good solution where morphological and syntactic similarities may be marked in an organized fashion for subsequent grammatical mapping and application. Since such a Core Grammar is yet to be developed for the genealogically/typologically related Indian languages, it is urgent that we should try to design and develop Core Grammars as well as bilingual/multilingual dictionaries for the Indian languages after analyzing BTCs produced in Indian languages.

12.8 Extraction of Translational Equivalents

The search for translation equivalents (TEs) in a BTC starts with the particular forms that express similar meanings or concepts in SL and TL. After these forms are identified in SL and TL texts, these are systematically stored in a separate lexical database. Normally, most of the BTCs yield a large amount of translation equivalent forms, which are potential to be used as alternative forms also. The factors that determine the choice of appropriate equivalent forms are measured on the basis of recurrent patterns of use of the terms. The equivalent forms found in a BTC may be

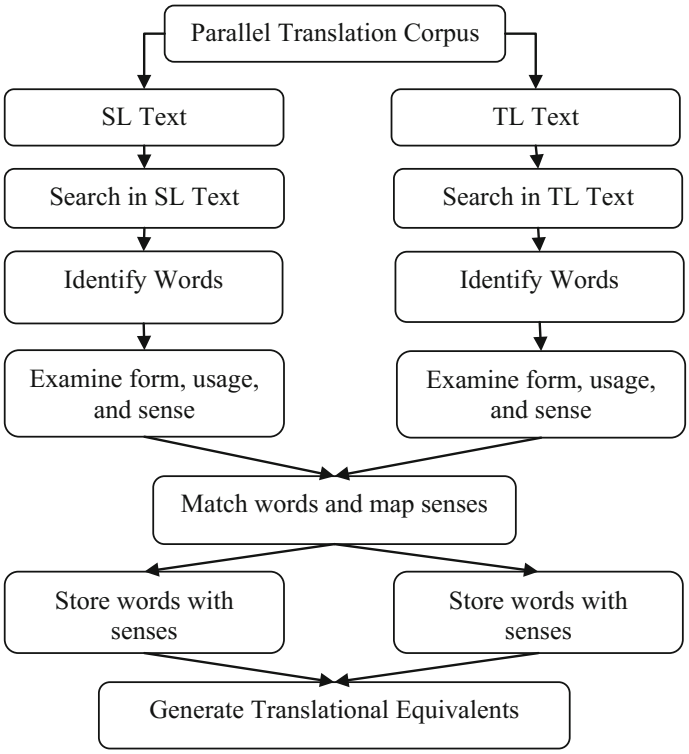

Fig. 12.4 Extraction of translational equivalents from a BTC

verified with the corpus of monolingual texts of SL and TL from which a BTC is produced.

It is to be understood that, even within two closely related languages, translation equivalent forms seldom mean the same thing in all possible contexts, since these are seldom used in the same syntactic and grammatical environments (Macken et al. 2013). Also, their semantic connotations and degree of formality may differ based on language-specific contexts. For example, a lemma in TL is rarely found to be an exact equivalent to a lemma of SL even though the terms are conceptually equivalent. The two-way translation is possible with proper names and scientific terms, but hardly with ordinary lexical items. It implies that a CBMT system will face tough problem due to differences in sense of ordinary words. We, therefore, require a high level of linguistic sophistication to yield better outputs through lexical comparison. Where there are few problems, as in the case of scientific and technical texts, a CBMT system may have a good result. But in case of imaginative texts, such a scheme will require more refined lexical selection process. The following diagram presents a scheme about how a list of translation equivalents is possible to generate from a BTC (Fig. 12.4).

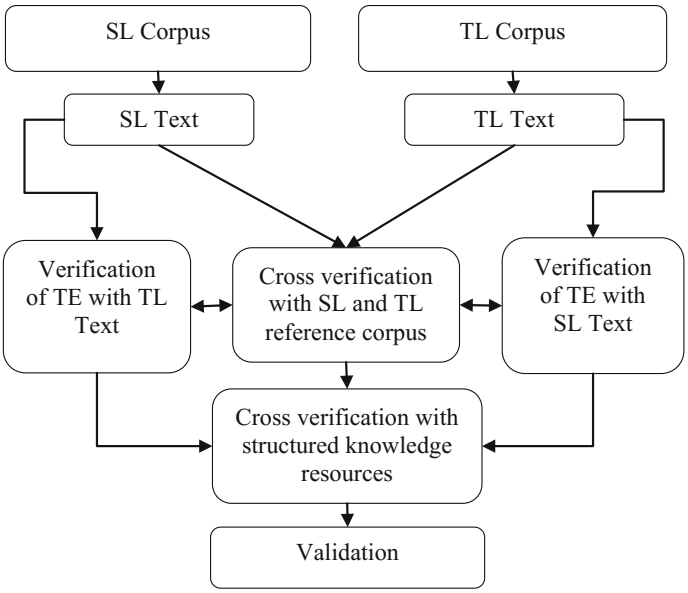

Fig. 12.5 Validation of translational equivalents in a BTC

Extraction of translation equivalents from a BTC can help the CBMT system designers to build on resources in multilingual areas, as a BTC empowers them to trace suitable translation equivalents. In general, a BTC enables all the language investigators to carry out the following tasks:

(a) Retrieve good translation equivalents including words, idioms, compounds, collocations, and phrases.
(b) Learn how language corpus help in producing translated texts that can exhibit 'naturalness' of a TL.
(c) Create new translation database, which may enable to translate correctly into languages on which one has limited access.
(d) Generate terminology data bank, for those languages which have no standard TermBank.

The process of extracting equivalent equivalents from a BTC and their subsequent authentication with reference to monolingual corpus and existing structured knowledge resources are shown in the following diagram (Fig. 12.5).

To find translational equivalents from a BTC, one has to apply various searching methods to trace the comparable units which are similar in meaning. Normally, some are larger and complex in form than simple words. Once these are identified, these are integrated into a translation platform, a database that facilitates translation more than customary translation memories. We may integrate findings from a BTC with bilingual dictionaries, multilingual TermBanks, and the databases of translational equivalents to enrich the CBMT knowledge base for the battles ahead.

For generating a good database of translational equivalents, the need for structural mappings between aligned BTC is widely recognized in CBMT method as well as in the other areas of translation. Till date, the flat-structured models, such as word-based equivalent models of the early 1990s or more recently, the phrase-based models are widely used for this purpose. In fact, the tree-structured mappings of phrases and sentences offer greater potentials validating relationships between the languages (Pala and Ganagashetty 2012). Using an aligned BTC we can work on inventing better translation models to develop better algorithms for a CBMT system. In general, the goals in this area are the followings:

(a) Designing robust syntax-based or tree-structured based models for a CBMT system,
(b) Designing machine learning techniques for inducing different structured translation models between the language pairs,
(c) Making an empirical analysis of an aligned BTC to study the level of adequacy and efficiency of the formalisms,
(d) Exploring the usefulness of parsed syntactic resources for a CBMT system between any two languages,
(e) Evaluating the scalability of structured translation methods to small or large corpus databases of language pairs,
(f) Utilizing information derived from structural formalisms in the area of speech translation, formal semantics and semantic parsing, paraphrasing and textual entailment, and information retrieval and extraction.

12.9 Generation of Terminology Data Bank

The coinage of appropriate technical and scientific terms for a CBMT system also requires semantic and functional analysis of the terms (Temmerman 2000). Here, the basic function is to search through TL texts and identify the appropriate terms that are considered equivalents or semi-equivalents for the 'foreign' ideas, items, or concepts found in SL texts. While doing this, one may need to keep various factors in mind regarding appropriateness, grammaticality, acceptability, and usability of the terms among TL users. But the most crucial factor is the feature of 'lexical generativity' of the terms so that new words are possible to generate by way of using various word formation processes available in TL (Wright and Budin 1997). A BTC has another important role in the selection of an appropriate term from a large list of terms collected or coined by the TL users for representing particular ideas, events, items, and concepts of SL. The process of developing TermBank is described in the previous chapter (Chap. 11).

It is noted that the recurrent practice of forming new technical terms goes to such an extreme that a CBMT system designer is quite confused to decide which terms (s)he should select over the other possible candidates. The debate also arises whether one should generate new terms or accept the terms of SL already absorbed in TL

by regular use. It is noted that some technical terms are adapted to such a level that it is impossible to trace their actual origin. In this case, a CBMT system designer has no problem, because these terms are already 'naturalized' in TL. For instance, at the time of developing a CBMT system for English-to-Bangla translation, one faces no problem in understanding the English terms like *computer, mobile, calculator, telephone, printer, tram, bus, cycle, taxi, rickshaw, train, machine, pen, pencil, paint, road, station, platform* because these are already naturalized in Bangla.

Moreover, their high frequency of use in various texts makes them a part of the modern Bangla vocabulary. There is no need to replace these terms at the time of translation. In this regard, a BTC is a highly valuable resource for the selection of appropriate terms expressing new ideas and concepts borrowed from SL to TL. Since a BTC is made with varieties of texts full of new technical terms, idioms, phrase, and expressions, it provides good-quality context-based examples to draw sensible inferences. Here, a BTC contributes in two ways.

(a) It helps to assemble technical terms, jargons, expressions, and phrases entered into TL with information of dates and domains of their entry and usage.
(b) It helps to find possible native coinages of the terms, jargons, expressions, and phrases along with respective domains and frequency of their use in the language.

These two factors largely help to determine relative acceptance or rejection of terms in TL. Examination of examples collected from a BTC also shows how a BTC can be useful in the selection of appropriate terms, phrases, and expressions which are essential elements in both manual and automatic translations.

12.10 Lexical Selection from TL

Selection of a particular lexical item from TL text to be used as the most appropriate equivalent for a lexical item from SL text is another complex task that requires the active and careful interference of an expert well versed in SL and TL (Condamines 2010). A corpus linguist has to select an appropriate term from a large collection of semantically similar terms available in TL text, which is nearest in sense and connotation to a term selected from SL text. A typical example of this kind is the use of verb based on the status of an agent (i.e., actor). In Bangla, for example, the use of a verb referring to 'act of eating' is highly restricted based on the status of the agent used as a subject of a sentence. Consider the following examples:

1(a) Bangla: bhagabān *prasād grahaṇ karen* (sub.: God)
English: God *eats.*
1(b) Bangla: mahāpuruṣ *bhojan karen* (sub.: great man)
English: Nobleman *eats.*
1(c) Bangla: bhadralok *āhār karen* (sub.: gentleman)
English: Gentleman *eats.*

1(d) Bangla: sādhāraṇ lok *khāy* (sub.: common man)
English: Common man *eats.*
1(e) Bangla: choṭalok *gele* (sub.: a person of lower class)
English: Layman *eats*.

When we analyze the examples presented above (1a–1e), we easily find that the selection of the appropriate equivalent term in Bangla for the English word *eating* is controlled by the status of the agent (i.e., subject) referred to in the sentences. If the person in the English text is a *divine man*, then the equivalent word in Bangla for the English word *eating* is *prasād grahaṇ karā* (1a), for a *great man* it is *bhojan karā* (1b), for a *gentleman* it is *āhār karā* (1c), for a *common man* it is *khāoyā* (1d), and for a *layman* who belongs to the lowest social strata marked by the scales of occupation, social prestige, and economy it is *gelā* (1e), although, in all senses, the core meaning of the terms is the same 'eating food.'

The task of a linguist is to find out the appropriate lexical items considering various sociolinguistic factors latently involved within the two languages considered for translation. These well-known examples show that lexical selection has to be taken care of generating sensible translation outputs. Although the problem is often handled carefully in case of human translation, it is often ignored in automatic translation practices. The best way to overcome the problem in a CBMT system is to list all the semantically similar forms in a separate list within a machine-readable dictionary (MRD) to be used later in translation. Such lexical databases are easy to extract from a BTC for both manual and machine translations.

Usually, there are several subject areas within a machine-readable dictionary (MRD) which is capable of providing relevant information about the selection of a subject area of a lexical item. Therefore, when processing a text, a CBMT designer has to select the subject area, which she feels appropriate to the type of text to be translated into TL. For instance, if an English text relating to politics is to be translated into Bangla, it makes sense to instruct the CBMT system to look for relevant terms of TL in the TermBank in the subject area 'politics' before it scours through the remaining part of the lexical database. For example, the term *jñāpan* 'inform' in Bangla has several translational equivalents based on the subject area of its use in SL text, as the following list (Table 12.2) shows.

The examples given above emphasize that we have to select the most appropriate term considering the subject area, which we are going to translate from SL text to TL text. Until such issues are addressed, an appropriate output cannot be achieved in TL. The generation of such lists of idioms and proverbs from a BTC can enhance the quality and robustness of a CBMT system because these databases are used to encompass figurative senses of terms found in SL and TL for stylistic representation and better comprehension of outputs.

Table 12.2 Lexical selection based on subject area

English	Bangla equivalents (selection is domain-based)
↓	↓
Inform	jānāno (giving general news or information)
Inform	raṭāno (spreading rumor or false information)
Inform	pracār (canvassing information for all)
Inform	bijñāpan (advertising an item or product, etc.)
Inform	sampracār (broadcast and telecast of news)
Inform	bijñapti (government circulars or notices for all)
Inform	ghoṣaṇā (an event of public reference)
Inform	dhārābhāṣya (running commentary on sports)
Inform	istehār (political campaign and propaganda)
Inform	pratibedan (reporting a piece of news in papers)
Inform	kīrtan (highlighting someone's achievement)

12.11 Dissolving Lexical Ambiguity

In a normal situation, a linguistic communication transfers information from a producer to a receiver by way of using language as a vehicle. Sometimes, however, this transfer of information is not free from ambiguity, one of the most common yet complex phenomena of a natural language. It is noted that ambiguity arises due to inadequacy in 'internal meaning' associated with a lexicon or due to the structure of an utterance used in an event of communication. Ambiguity may be classified into three broad types:

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

At the level of lexical ambiguity, a speaker or a writer uses a single word to refer more than one sense, event, or concept. This creates a problem on the part of the listeners in capturing the actual intended meaning of the word. The problem intensifies when the language of a speaker differs from that of a listener. Since a CBMT system tries to capture the mental representation of a speaker, it is limited to words and sentences used by a speaker.

To overcome this problem, one needs to map the source lexicon with equivalent target lexicon. This mapping turns into an appropriate frame in particular contexts and situations of text representation. In some situations, a TL may not have an equivalent lexical item which is fit to represent the actual sense of a term used in SL. In such a case, we have to either depend on a cluster of words (e.g., *multiword units, compounds, idioms, phrases,* and *clauses*) or add explanatory addendum to deal with the situations.

For dissolving lexical ambiguities, an easier solution is to find out the methods for locating contexts of use of words as well as analyze the contextual profiles of lexical items. Recent experiments with corpus reveal that lexical ambiguity is caused due to multiple readings of words. These readings may differ in terms of subcategorization and the selection of features, syntactic and semantic properties, and features such as tense, modality, case, number, the possibility of idiomatic reading. For instance, Bangla words like *māthā* 'head,' *kathā* 'word,' *lok* 'person,' *path* 'way,' *karā* 'to do,' *khāoyā* 'to eat,' *kāṭā* 'to cut,' *k˜ācā* 'raw,' *pākā* 'ripe,' *bhāla* 'good,' *khārāp* 'bad,' *kāche* 'near,' *upare* 'above' are associated with multiple readings due to variation of context of use, which in return triggers their multiple lexical and figurative senses (discussed in Chap. 10).

As the supporters of a CBMT system we argue that to overcome the problem of lexical ambiguity, we should collect and analyze a large number of ambiguous forms, which occur in SL and TL texts and overtly represent them in an MRD to achieve accuracy in translation. If possible, we should analyze all ambiguous forms with knowledge obtained from BTC as well as with semantic information stored in structured knowledge resources.

On the other hand, one can argue that such a task of information acquisition from BTC is neither realistic nor feasible. One can also argue that a general translation process does not necessarily require a full understanding of a text. Ambiguities may be preserved during a translation process, and they should be presented to the users (i.e., readers) for resolution. Taking cues from domain-specific translation output, one can argue that deep semantic analysis of lexical items is not always needed for translation. For instance, English word *head* may be translated into Bangla as *māthā*, no matter in which sense it is used in SL text. Therefore, it is better to opt for a simple word analysis scheme and use a more direct 'SL-TL substitution' method in place of deep semantic analysis of ambiguous lexical items.

We agree that in certain contexts, it is possible and necessary to ignore certain lexical ambiguities with a hope that the same ambiguities should be carried over in translation to TL text. This is particularly useful in those translation systems that aim at dealing with only a pair of closely related languages within a highly restricted domain. However, since an understanding of lexical ambiguities is meant to produce a non-ambiguous representation in TL, it cannot be ignored in the translation of texts of both general and special domains.

12.12 Grammatical Mapping

The type of transformation we refer here is known as 'grammatical mapping' in translation. Here, the words of SL text are 'grammatically mapped' with the words of TL text to obtain a meaningful translation. There are various schemes for mapping used in a CBMT system (e.g., *lexical mapping, morphological mapping, grammatical mapping, phrasal mapping, clausal mapping,* and *syntactic mapping*). The most

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

Fig. 12.6 Grammatical mapping between English and Bengali

common form of grammatical mapping is related to verb forms within the two languages considered for translation.

The event of grammatical mapping becomes relevant in the context of a CBMT system between two languages, which are different in lexical ordering in case of sentence formation. In the present context when we talk about a CBMT system from English to Bangla, it is optimized in proportion, since English has SVO structure (e.g., $He_{[Sub.]}$ $eats_{[Fv]}$ $rice_{[Obj.]}$) in sentence formation and Bangla has SOV structure (e.g., $se_{[Sub.]}$ $bhāt_{[Obj.]}$ $khāy_{[Fv]}$) in the same framework. Therefore, a kind of grammatical mapping and reordering of the lexical items is needed for producing acceptable output in Bangla. Let us consider the examples given below (6a–6b) and the mapping diagram (Fig. 12.6).

6a. English: All his efforts ended in smoke.
6b. Bengali: tār samasta ceṣṭā byārtha hala.

It shows that to achieve accurate output with acceptable word order in TL text, the system has to map the words in TL text with the words used in a sentence of SL text in the following manner:

Lexical Mapping:

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

Even then it is noted that lexical mapping is not enough for obtaining proper translation outputs. The input sentence of SL text (English) contains an idiomatic expression (i.e., *ended in smoke*), which requires some extralinguistic knowledge to find a similar idiomatic expression from TL text (Bangla) to achieve accuracy in output. Therefore, it has to employ appropriate extralinguistic knowledge for selecting the most appropriate equivalent idiomatic expression from TL text (Bangla) in the following manner:

Extralinguistic Information:

English: [d–e–f] (an idiomatic expression),

Table 12.3 Mapping of postposition between English and Bangla

No.	English	Bangla
(1)	*In* hands	hāte (<hāt$_{[N]}$ + −e$_{[loc_case]}$)
(2)	*With* person	loker (<lok$_{[N]}$ + −er$_{[gen_case]}$) + sange$_{[pp]}$)
(3)	*By* mistake	bhulbaśata (<bhul$_{[N]}$ + baśata$_{[ADV]}$)
(4)	*In* house	ghare (<ghar$_{[N]}$ + −e$_{[loc_case]}$)
(5)	*In* house	gharer madhye (<ghar$_{[N]}$ + −er$_{[gen_case]}$ + madhye$_{[pp]}$)
(6)	*At* night	rāte (<rāt$_{[N]}$ + −e$_{[loc_case]}$)

Bangla: [7 (<4–5–6)] (similar translation equivalent).

A CBMT system needs to be provided with information that *ended in smoke* in SL text is to be translated as *byārtha hala* in TL text when the expression is used in an idiomatic sense. After the selection of appropriate and equivalent idiomatic expression from TL text, the system is fixed in a position to show that the output sentence is grammatically mapped to such an extent that the intended sense of the input sentence is maximally represented. After this comes the stage of sequential ordering of words in the sentence of TL text so that the output sentence becomes grammatically acceptable in TL. For this, the following information is handy:
Sentential Information:

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

After proper application of several linguistic strategies like lexical mapping, the selection of appropriate idiomatic expression (if any) and sequential ordering, one finally gets *tār samasta ceṣṭā byārtha hala* as an acceptable translation in TL. In essence, such grammatical mapping from one structure to another is used to produce suitable translations that may be accepted as 'normal' constructions in TL.

In this task, therefore, the analysis of sentence structures of translated text is highly necessary. From a translated text, it is possible to map the sequence of word order (at the linear level) between SL texts and TL texts to yield valuable information about the structure of NPs, APs, VPs, PPs, and other properties used in the languages considered for bidirectional translation.

The grammatical mapping highlights lexical interface that underlies surface structures of sentences and the nature of lexical dependency underlying surface constructions. For instance, in case of translation of prepositions (e.g., *at, up, by, in, of, with, for*) used in English, one has to decide whether postpositions or case markers have to be used in Bangla translation. For elucidation, consider the following examples given below (Table 12.3).

The above examples (Table 12.3) show that in English, prepositions occur before nouns to evoke case relation (1, 4, and 6), adverbial sense (3), and postpositional sense (2 and 4). All these senses may be achieved in Bangla by way of using either case markers (1, 4, and 6), postpositions (2 and 3) or both case markers and postpositions

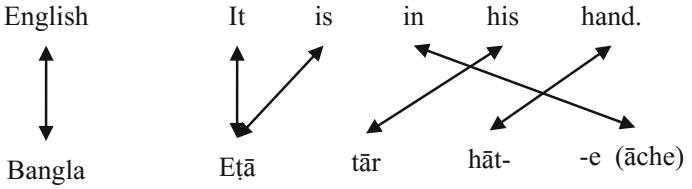

Fig. 12.7 Position of postposition with respect to content words

(5). Moreover, it also provides information about their position with regard to the content words with which these functional words are attached to (Fig. 12.7).

From the examples and analysis presented above, it is clear that the task of proper grammatical mapping is an essential component of a CBMT system, which we cannot ignore if we really want to achieve success in this area.

12.13 Other Issues in CBMT

There are four major types of issue that are involved in the development of a CBMT system such as the followings:

(1) **Lexical Issues**: morphological analysis, bilingual dictionary, translational equivalents, lexical mismatch, lexical ambiguity, lexical gap, lexical collocation, the mapping between single-word unit and multiword unit, named entity mapping, etc.,
(2) **Syntactic Issues**: lexical reordering, structural ambiguity, anaphoric ambiguity, sentence splitting, sentence joining, grammatical agreement, sentence alignment, etc.,
(3) **Semantic Issues**: semantic analysis to identify the meaning of various linguistic units (i.e., words, idioms, phrases, etc.), and the ambiguities involved therein,
(4) **Sociocultural Issues**: idiomatic expressions, proverbial expressions, jargons, discourse knowledge, cultural knowledge, semiotic information, ecolinguistic factors, extralinguistic knowledge, etc.

12.14 The Modular System

We understand that the analysis of BTC can help to optimize mapping between the two equivalent constructions in order to obtain a better translation. Usually, it involves associating equivalent constructions (e.g., *multiword units, idioms, phrases, clauses, larger syntactic structures*) that are endowed with typical formal structures at the time of corpus analysis. However, the basic purpose of this process is to allow pairing mechanism to be divided into three parts in a systematic way:

(1) The identification of potential linguistic units, which may be grammatically associated in the two corpora,
(2) The formalization of structures of associable units by way of using sets of morphosyntactic and lexicosemantic tags,
(3) The determination of the probability of the proposed structures comparing these forms with effective databases collected from manually translated texts.

By subdividing the entire process into three phases, a relatively simple translation module may be produced to determine the units likely to correlate with theoretical analysis of translations observed in a BTC. One possible solution to make it easier is to develop analysis methods based on data stored in a 'training corpus.' Such a method, based on model training, may depend on the amount of linguistic information available *a priori*, i.e., on the syntactic rules previously developed by humans (Somers 2008). It should be, however, noted that it is not necessary to analyze all sentences used in a BTC to find out the rules of respective languages. Analysis of a set of type constructions rather than the full sets of tokens will suffice and serve the initial purpose, because of the following reasons:

(1) Within each language, there are linguistic constructions, which are identical in form and composition to others. That means, an NP may correspond structurally to some NPs within a text. This is more or less true to both SL and TL.
(2) The sequence and interrelationships between the units in TL text may be the same with those in SL text if a BTC is developed from two closely related sister languages.
(3) There are certain fixed reference points, which easily mark out texts and mark identification of translation units. Such reference points include numbers, dates, proper nouns, titles, paragraphs, sections of the two texts.

Based on analysis of equivalent forms obtained from a BTC, one can find out three types of grammatical matching:

(1) Examples of 'strong match' where a number of words, their order, and their meanings are same,
(2) Examples of 'approximate match' where a number of words and their meanings are same, but not the order in which they appear in texts,
(3) Examples of 'weak match' where order and number of words are different, but their dictionary meanings are same.

At the time of translating texts from English to Bangla, most of the grammatical mappings will belong to the class of 'weak match,' since the languages belong to two different typologies (English SVO type while Bangla SOV type). In such a situation, alignment of BTC of the two languages should not rely only on the syntactic structure of respective texts but should have greater scopes for semantic anchor points. We argue that if 70 percent words in a sentence of SL text semantically correspond to at least 70 percent words in a sentence of TL text, then we may claim that the sentences have semantic equivalency to have a translational relationship. Reliability of such a CBMT system may be measured by an intermediate alignment stage at the paragraph or at the sentence level.

To ensure and achieve greater reliability, one may use the 'regressive alignment technique' that starts with large text units (e.g., *chapters and paragraphs*) to proceed toward small units (e.g., *sentences, phrases,* and *words*) on a BTC. The application of this method on a BTC (focusing on large to small units) will enable one to verify the following hypotheses.

(1) Two chapters have translation relationship if at least 70% paragraphs correspond to each other.
(2) Two paragraphs have translation relationship if at least 70% sentences correspond to each other.
(3) Two sentences have translation relationship if at least 70% words correspond to each other.
(4) Two words have a translational relationship if at least one of their meanings (stored in the bilingual dictionary) corresponds to each other.

An insightful combination of a CBMT system with an SBMT system can make it possible to fine-tune the alignment process of a BTC to enhance text processing and information collection. However, it requires identification and formalization of 'translation units' and utilization of bilingual dictionaries and MRDs. So, there is no need for exhaustive morphosyntactic tagging of each text, since a machine can do it with statistical support to find out equivalent forms by comparing texts in a BTC that exhibits translational relations. However, to ensure quality performance of a CBMT system one has to take care of the following things.

(a) The standard of a BTC should be high. Aligned bilingual texts may pose a problem if the quality of texts is poor or if these are not put under the strict vigilance of linguists.
(b) The quality and size of a bilingual dictionary should be large. A dictionary is an indispensable resource in terms of providing proper grammatical information. Besides, it should have a provision to integrate unknown words found in a BTC.
(c) The robustness of a CBMT system and the quality of translation will depend on the volume of training data available.
(d) The level of accuracy in translation outputs will rely heavily on the levels of synchronization between the texts of a BTC.

With above resources and methodologies, a CBMT system is bound to be robust, provided a long training phase on different types of text is carried out. Once the training phase is completed, information stored in translation memory will be activated to yield all types of translation solutions exclusive to human experts. However, to achieve a greater level of success, a dose of artificial intelligence has to be integrated into a CBMT system.

12.15 Conclusion

A CBMT system is an ideal field for comprehensive evaluation of various computational theories of language and for the development and testing of a wider range of linguistic phenomena abundant in natural languages. The availability of BTC makes a significant contribution to enhancing robustness and capability of a CBMT system. Success in this method directs us to supplement traditional rule-based approaches since information obtained from analysis of a BTC minimizes the distance between SL and TL (Teubert 2000).

A BTC performs a two-way role. It provides inputs for developing an MT system, and it supplies texts for evaluating an MT system. Thus, a BTC makes a significant contribution toward enhancing the capability of an MT system. We know that a CBMT system is successful in many domain-specific translations with controlled language databases where all kinds of syntactic, lexical, and idiomatic ambiguities are dissolved. It narrows down the gulf of mutual intelligibility to enhance the level of translatability between the two languages used in translation. What it confirms is that if we are interested to develop a good CBMT system for the Indian languages, we should develop good-quality BTC among the Indian languages.

References

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

Baker, M. 1993. Corpus linguistics and translation studies: Implications and applications. In *Text and Technology: In Honour of John Sinclair*, ed. M. Baker, F. Gill, and E. Tognini-Bonelli, 233–250. Philadelphia: John Benjamins.

Baker, M. 1996. Corpus-based translation studies: The challenges that lie ahead. In: *Terminology, LSP, and Translation: Studies in language engineering in honour of Juan C. Sager.*, ed. Somers, H. Translation Library 18, 175–186. Amsterdam: John Benjamins'.

Brown, P., J. Cocke, S.D. Pietra, F. Jelinek, R.L. Mercer, and P.S. Rosin. 1990. A Statistical approach to language translation. *Computational Linguistics* 16 (1): 79–85.

Brown, P.F., S.D. Pietra, and R.L. Mercer. 1993. Statistical machine translation. *Computational Linguistics* 19 (2): 263–312.

Castillo, J.J. 2010. Using machine translation systems to expand a corpus in textual entailment. In *Proceedings of the 7th International Conference on Advances in Natural Language Processing*. New York, US: Springer, 97–102.

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

Condamines, A. 2010. Variations in terminology: Application to the management of risks related to language use in the workplace. *Terminology* 16 (1): 30–50.

Dash, N.S. 2005. Role of context in word sense disambiguation. *Indian Linguistics* 66 (1–4): 159–175.

Dash, N.S. 2016. Culling scientific and technical terms (STTs) from text corpora for compiling termbank in Bangla. *Research Cell: An International Journal of Engineering Sciences* 21: 107–122.
Dash, N.S., and S. Arulmozi. 2016. Generating parallel translation corpora in indian languages: cultivating bilingual texts for cross-lingual fertilization. *Translation Today* 10 (1): 84–118.
Dietzel, S. 2009. *Example-based Machine Translation*. Berlin: Springer.
Furuse, O., and H. Lida. 1992. An Example-based Method for Transfer-driven Machine Translation. In *Proceedings of the MTI-92, Montreal, Canada,* 139–150.
Jones, D. 1992. Non-hybrid Example-based Machine Translation Architectures. In *Proceedings of the MTI-92,* Montreal, Canada, 163–171.
Kay, M., and M. Röscheisen. 1993. Text-translation alignment. *Computational Linguistics* 19 (1): 13–27.
Koehn, P. 2005. Europarl: a parallel corpus for statistical machine translation. In *Proceedings of MT Summit X*, Phuket, Thailand, 79–97.
Koehn, P. 2010. *Statistical Machine Translation*. Cambridge: Cambridge University Press.
Macken, L., E. Lefever, and V. Hoste. 2013. Bilingual terminology extraction from parallel corpora using chunk-based alignment. *Terminology* 19 (1): 1–30.
McLean, I. 1992. Example-based machine translation using connectionist matching. In *Proceedings of the MTI-92*. *Montreal, Canada*, 35–43.
Pala, K., and S.V. Ganagashetty. 2012. Challenges and opportunities in automatically building bilingual lexicon from web corpus. *Interdisciplinary Journal of Linguistics* 5 (1–2): 169–184.
Sanderson, M., and W.B. Croft. 2012. The History of information retrieval research. *Proceedings of the IEEE* 100: 1444–1451.
Somers, H. 1999. Example-based machine translation. *Machine Translation* 14 (2): 113–157.
Somers, H. 2008. Corpora and machine translation. In *Corpus Linguistics: An International Handbook*, ed. Lüdeling, A., and M. Kytö, 1175–1196. Berlin: Mouton de Gruyter.
Su, K.Y., and J.S. Chang. 1992. Why corpus-based statistics-oriented machine translation. In *The Proceedings of the MTI-92*, Montreal, Canada, pp. 249–262.
Temmerman, R. 2000. *Towards New Ways of Terminology Description: The Socio-Cognitive Approach*, 26. London: John Benjamins.
Teubert, W. 2000. Corpus linguistics—A partisan view. *International Journal of Corpus Linguistics.* 4 (1): 1–16.
Teubert, W. 2002. The role of parallel corpora in translation and multilingual lexicography. In *Lexis in Contrast: Corpus-based Approaches*, ed. B. Altenberg and S. Granger, 189–214. Amsterdam: John Benjamins.
Vandeghinste, V. 2007. Removing the distinction between a translation memory, a bilingual dictionary, and a parallel corpus. In *Proceedings of Translation and the Computer 29*, ASLIB, London, UK.
Winograd, T. 1983. *Language as a Cognitive Process*, vol. I. Mass: Addison-Wesley.
Wright, S.E., and G. Budin. 1997. *Handbook of Terminology Management, Basic Aspects of Terminology Management*, vol. 1, 370. Amsterdam: John Benjamins.

Web Links

http://www.dfki.de/~hansu/LT.pdf.
http://www.dfki.de/~hansu/HLT-Survey.pdf.
http://clt.gu.se/.
https://en.wikipedia.org/wiki/Linguistic_Issues_in_Language_Technology.
https://www.lti.cs.cmu.edu/.
http://hltcoe.jhu.edu/.

https://www.mq.edu.au/.
http://nlp.stanford.edu/IR-book/pdf/08eval.pdf.
http://singhal.info/ieee2001.pdf.
http://nlp.stanford.edu/IR-book/pdf/01bool.pdf.
http://unicode.org/standard/WhatIsUnicode.html.

Chapter 13
Corpus and Some Other Domains

Abstract Language corpus is now accepted as one of the primary resources in several branches of application-oriented and description-based linguistics. In all these branches, corpus is directly and indirectly used for description, analysis, and application of various elements and properties of a language. This trend of using corpus as a resource actually reflects on the ideological shift in the approach of the language investigators and applicators in recent years. The ready availability of corpus has made us realize that we do not need to depend on our intuitive linguistic expertise to establish our claims. Rather there is a great scope for us to extract data and information from a corpus for the same purpose. This alternative method of language study has inspired us to depend on language data faithfully obtained from real-life situations rather than depending on our intuitive speculation. Keeping this phenomenon in view, in this chapter, we shall try to show that the utility of language corpus is no more confined within a few areas of linguistics and language technology. Rather it is being used in many old and new branches of linguistics to make these fields more useful, informative, and insightful. To substantiate our argument, we shall describe the use of corpus in some important domains of linguistics, namely lexicology, lexical semantics, sociolinguistics, psycholinguistics, and stylistics.

Keywords Corpus · Linguistics · Lexicology · Lexical semantics
Sociolinguistics · Psycholinguistics · Stylistics

13.1 Introduction

The introduction of corpus has induced a new lease of life to many disciplines. The advancement made in computer science has given us a new scope to produce several large-sized language corpora in digital form and use these in different areas of language research and development. This new way of looking into language data has added a new dimension to the traditional scheme of linguistic studies. This has been possible due to advanced computer technology that contributes heavily to corpus linguistics by way of supplying useful tools and techniques for gathering evidence

N. S. Dash and L. Ramamoorthy, *Utility and Application of Language Corpora*,
https://doi.org/10.1007/978-981-13-1801-6_13

of actual language use and analyzing these from new angles and perspectives. In our view, the use of language corpus has contributed in five major ways to the field of linguistics in general:

(a) It has given descriptive linguistics a new opportunity to describe a language or some of its features with new sets of data, information, and examples.
(b) It has empowered theoreticians to verify whether the age-old theories and models about the language and language use are worth pursuing.
(c) It has given applied linguists an opportunity for direct utilization of language data in different works of applied linguistics and language technology.
(d) It has given new opportunity to the social scientists to use the varied and large amount of language data to form new observation about the speakers and the speech community in light of new types of language data and information.
(e) It has given a new lease o life to linguistics which was suffering for years due to the limited scope of its direction, diversion, and application.

There was a time when some linguists wanted to terminate of all kinds of empirical linguistic research by emphasizing only on generative methods. During that time, some serious linguists working in language teaching, dictionary making, language acquisition, speech analysis, text analysis, translation, dialogue analysis, etc., started using data from language corpora and presented new results to prove that the use of a corpus in linguistics is quite beneficial, because a corpus produces something new, which has not been possible to see by any other method (Francis 1982).

It had a far-reaching impact on the world of linguistics. With the change in attitude, corpora are used in other fields of linguistics. As a result, many new advantages, which linguists wanted to have but could not get any chance to avail, are now made available to them. Probably, the modern computational facilities have been the reasons behind the exodus of a large number of linguists from the state of intuitive infancy to the adulthood of scientific orientation of language study.

The quantum improvement in storage, processing, management, analysis, utilization, and customization of a large amount of language data by computer has made a lasting impact in diversified growth and expansion of language corpus and making it an indispensable resource for many classified works of descriptive and applied linguistics. The computer has provided facilities like massive storage, impressive text processing speed, elegant data management, fast data access interface, accurate text analysis mechanism, and quick result generation facilities. All these advantages have become tremendously beneficial for many established branches of linguistics, such as lexicography, language teaching, sociolinguistics, psycholinguistics, grammar writing, stylistics, where the use of computer and language corpus, even a few decades ago, was a fanciful daydream (Fig. 13.1).

For instance, when we look at the some recently published English dictionaries (e.g., *Random House Unabridged Dictionary* (1993), *Longman Dictionary of Contemporary English* (1995), *Oxford Advanced Learner's Dictionary* (1995), *Collins COBILD English Dictionary* (1995), *Cambridge International English Dictionary* (1995), *Encarta World English Dictionary* (1999)), we find that all these dictionaries

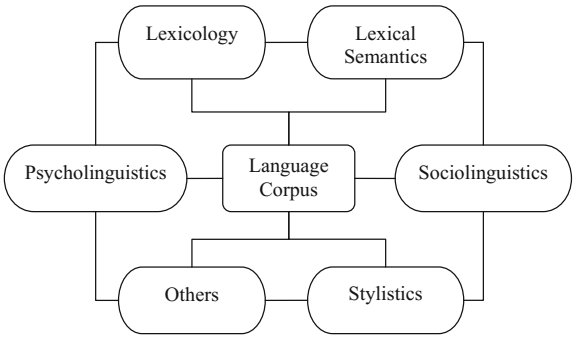

Fig. 13.1 Corpus and some distant daughters of linguistics

are made with data and examples taken from corpus, and as a result of this, these are more informative, more exhaustive, more authentic, more reliable, more representative, more useful, and more updated than our handmade printed dictionaries that are usually compiled manually with lexical data obtained from secondary sources. Most of these dictionaries contain detail corpus-based information about meaning and usage of headwords along with elaborate information about their lexicological identity, contextual use, polysemous nature, and collocational network, besides regular etymological, phonological, morphological, and semantic information that is normally found to be present in standard printed dictionaries. In essence, a language corpus has succeeded to establish its application relevance in most of the established branches of linguistics because it is able to provide reliable and diversified information hardly obtainable from intuition. The impact has been so strong that now people are reluctant to work in any branch of linguistics without reference to a corpus.

Trailing the track of discussion presented in earlier chapters, in this chapter, we want to focus on the utilization of corpus in some major branches of linguistics. We shall try to show how a corpus can be utilized in each branch of linguistics as a resource to verify earlier observation, extract new linguistic evidence, define new linguistic rules, construct new models, and formulate new theories. Since we cannot describe all the fields in this chapter, we focus on a few branches of linguistics like lexicology, lexical semantics, sociolinguistics, psycholinguistics, and stylistics. In Sect. 13.2, we shall focus on the interface underlying between a corpus and lexicology; in Sect. 13.3, we shall address how information collected from a corpus can be used in the area of lexical semantics; in Sect. 13.4 we shall show how corpus can contribute for the betterment of sociolinguistics; in Sect. 13.5, we shall discuss psycholinguistics can be enriched with data and information taken from a corpus; and in Sect. 13.6, we shall show how a corpus can become quite useful in the study of stylistics.

13.2 Corpus and Lexicology

The discipline of lexicology is an established branch of linguistics that has received an adequate amount of attention from the scholars working in semantics, lexicography, semasiology, semiotics, culture, onomasiology, ethnology, and others. At present, this branch sustains and grows with direct reference and use of data and information collected from the general and historical corpus (Vera 2002). Lexicology covers a wide range of interests relating to lexical investigation and analysis. Generally, the study of lexicology includes the following issues:

(1) Reconstruction of original meanings of words, which have undergone diachronic semantic changes with the change in time,
(2) Study of the patterns of lexical variation and change across a particular timescale,
(3) Evolution of the vocabulary of a language over the centuries,
(4) Study of neologism and lexical loss undergoing within a language,
(5) Study of lexical borrowing and derivation over a particular time span in a language,
(6) Structural and etymological analysis of lexical items used in a language.

With a close conceptual network existing between semantics, syntax, discourse, lexicography, pragmatics, and semiotics, the domain of lexicology fabricates an interface for cognitive interpretation of words and their senses. Although this kind of lexical analysis is yet to flourish in many languages, there are some recent works in English and other languages, which are mostly based on analysis of corpus data of various types (Coleman and Kay 2000). Such works have opened new perspectives on language investigation to show how a corpus can contribute to the growth and expansion of our knowledge of a language and its speakers. Some of the studies are briefly discussed in the following paragraphs.

The direction of semantic change of some abstract nouns of English is studied with reference to the *Helsinki Corpus of English Texts*, the *Michigan Early Modern English Materials Corpus*, and the *Oxford English Dictionary* (Alanko 2000). The study specifies the mechanisms involved in the nature of semantic change of nouns to highlight the cognitive processes involved in it. A chronological analysis of the corpus data makes it evident that the study of semantic change of words needs tools such as prototypes and materials such as historical corpora to document reliably the directions in semantic change and subjectification of abstract nouns in conceptualization. The analysis of corpus data clearly shows that the basic meanings do not disappear from a semantic field in general, although they may be lost in case of individual words. The patterns of semantic change are repeated in the same manner in the formation of new meanings of words inside the prototypical center so that the picture looks the same over the generations.

The *Collection of English Language Corpora* (1999) is analyzed to identify the reasons to explain the phenomenon of word loss and semantic change of words in the Middle English language (Cabanillas and Martínez 2002). The study explores

the ways by which the newly introduced lexical items can influence the recipient language and affect the native words. The corpus-based account of descriptive meanings of terms as well as the process of metaphorization of meanings of words reveals the effect of metaphor on some lexical items in the process of their semantic development in the language. With evidence obtained from a corpus, the study also reflects on the notable change in the original meaning of some lexical items caused due to use of metaphors at different points in time in the process of semantic change that operates in a language across spatiotemporal dimensions.

The *Toronto Corpus of Middle English*, the *Historical Thesaurus of English*, and the *Oxford English Dictionary* are used to as a corpus database to reconstruct the literal, metaphorical, and metonymical senses expressed by different lexical fields of some common English words (Gevaert 2002). A comparative study of the results obtained from analysis of corpus data shows that the basic conceptual fields of words undergo change under the influence of foreign concepts to redress the balance of ideas. The study points out that evolution and interaction of words may be measured by an innovative way which combines historical, cognitive, and prototype semantic approach based on quantitative analysis of historical corpus data.

The *Helsinki Corpus of Middle English* and *A Representative Corpus of Historical English Registers* are analyzed in historical perspective to trace the changes in word formation patterns as well as to address the general dynamics of word formation in English (Cowie and Puffer 2002). The analysis of the methods of productivity in word formation, as a qualitative–quantitative and diachronic process, reveals that there are various processes in word formation, which often undergo changes at different points in time. Due to this reason, the phenomenon of lexical productivity of a language should never be considered as a purely theoretical construct. On the contrary, it should be considered as a measurable feature that operates over a longer period of time, and therefore, it should be studied with reference to historical and diachronic corpora.

The *Helsinki Corpus of Modern English*, the *Lampeter Corpus of English*, the *Corpus of Early Modern English Correspondence Samples*, the *Michigan Early Modern English Materials*, and the *Corpus of Middle English Prose and Verse* are analyzed in a recent study to examine the importance of rhetorical purpose and context in semantic change of lexical items (Lewis 2002). The study shows that various scalar qualifiers with the representational function may be developed for analyzing the polysemous expressions to serve both the epistemic and evaluative functions of the lexical items. It also shows that subjectification of meaning arises from the lexical items due to their use in regular rhetorical patterns that lead to their semantic shifts via local analogies, which eventually force the lexical items to extend their usages in newer domains. The quantitative analysis of expressions from the corpora shows that co-occurrence of words with particular rhetorical patterns usually generates new polysemy to acquire new information structure.

The *Helsinki Corpus of Early Modern English*, the *Brown Corpus*, and the *Lancaster-Oslo-Bergen (LOB) Corpus* are analyzed to explain the change in meaning of the English word LOVE over the last 500 years (Tissari 2000). A multidirectional analysis of the data obtained from these corpora identifies five different conceptual domains for the word: family love, friendship, sexual love, religious love, and love

for things. Although there are doubts about the use of the word in certain contexts, reference to the participants involved in domain identification suggests that the word is versatile in conceptual overlapping in its function of sense denotation. Therefore, more than one category is denoted by the word in its sense, although contextual information differs with regard to the participants. The numerical analysis of corpus data, on the other hand, suggests that relative frequency of the domains of the word changes over the years. While sexual love remains the most dominant over the centuries, family love and friendship have become less frequent, and love for things and religious love remain almost constant over the centuries.

Some randomly sampled English historical corpora are analyzed to trace the semantic evolution of troponyms of 'look at' in the field of visual perception (Poch and Clavera 2002). The list of regular troponyms for 'look at' in English includes *stare, gaze, gape, gawp, gawk, goggle, glare, glimpse, glance, peek, peep, peer, squint, leer, gloat,* and *ogle*, where each form has a distinct denotative sense based on its context of use in a piece of text. The study highlights diverse semantic domains from which these verb forms are originated and focuses on the factors that motivate transfer of senses from one domain to another. Thus, from a purely cognitive perspective, the study shows how the present state of visual perception is reached. A simple diachronic survey of the lexical database collected from the corpora shows that most of the verbs have entered into English lexicon in the middle and modern age since only a few of them are present in Old English vocabulary. Although the origin of some of the verbs is obscured, it is noteworthy that their first documented senses are not often related to visual perception. The most striking observation, however, is that not only these verbs but also those words that are connected with visual perception reflect on the fact that eyes, apart from their basic functions of seeing or looking at, can also express various mental states like feelings, emotions, anxieties, and attitudes (Poch and Clavera 2002: 571).

Such studies reveal that reference to the history of corpus-based studies in lexicology has an explicit significance. It implies that both historical and diachronic corpora are available for excessive use in lexicology for making a significant contribution in historical semantics and lexicography (Hundt 1997). The process of rediscovery of lexical meaning in historical linguistics is benefited greatly from the generation of a wide range of diachronic corpora and corpus-based study materials (e.g., diachronic dictionaries and historical thesauri). These materials have allowed us to fine-tune our analysis on the evolution of the meaning of words of a language across time. Although the method of corpus-based study of word meanings is accepted as an important area of study in lexicology and historical linguistics in English and other languages (Hofland and Johansson 1982), it is yet to start its journey in the Indian languages.

13.3 Corpus and Lexical Semantics

Recent trends of a corpus-based approach to language study have contributed toward the establishment of an object-oriented approach to the semantic study of linguistic items and text segments. The basic view of the method is that the actual meaning of lexical items may be derived from the contexts in which they occur (Schütze 1997: 142). Although the meaning of the lexical items combines the history of their previous occurrences with the meanings of the parts they are made of with, finer shades of meaning (e.g., denotative, figurative, metaphoric, stylistic) are available from the contexts of their actual occurrence (Teubert 2000). Generally, these finer shades of meaning are condensed and paraphrased into a text that describes the meaning of the lexical items.

Meanings of lexical items are traditionally described with regard to our own intuition or knowledge of a language. However, information derived from analysis of a corpus reveals that semantic distinctions of words are associated with several characteristically observable contexts marked with figurative, morphological, syntactic, prosodic, and idiomatic frames. In a similar fashion, the meanings of larger constructs like compounds, multiwords, collocations, and phrases are linked with the contexts of their occurrence. Therefore, decipherment of contextual meanings of these forms needs carefully extracted contextual information from a corpus for proper semantic analysis and understanding. Within last few years, several empirical experiments are carried out to show how information obtained from a corpus may be used to provide objective criteria for assigning meaning to various linguistic items (Mindt 1991). This entails that consideration of environments of occurrence of various linguistic entities will provide objective interpretation to build up their semantic distinctions.

The importance of a language corpus, in lexical semantics, has been acknowledged in establishing the notions of 'fuzzy meaning' (i.e., semantic indeterminacy) and 'semantic gradiance' in understanding the meaning of words (Leech et al. 1994). In lexical semantics, functional categories of lexical items are usually constant. That means a lexical item either belongs to a particular category, or it does not. However, experiments carried on categorization suggest that functional categories of words have 'fuzzy' boundaries rather than one-dimensional constant frames. It implies that it is not the question whether a word belongs to one category or the other. It is important to know how often a word falls into one category as opposed to other. By looking at examples collected from a corpus, it is possible to show that the concept of 'fuzzy meaning' suits better for words since there is no clear boundary existing among the categories of words. Indeed there are phenomena of true 'semantic gradiance,' which are connected with 'frequency of inclusion' of words to a particular functional category based on their use in specific contexts.

The reference to a corpus can help us understand the nature of polysemy by which a word denotes multiple senses triggered from variations in contextual occurrence (Ravin and Leacock 2000; Bouillon and Busa 2001). In recent experiments, it is observed that the number of senses of a polysemous word that show up in a corpus exceeds the number of senses cited in a handmade dictionary (Fillmore and Atkins

2000). Therefore, it is sensible to refer to a corpus if we really want to understand all the sense and usage variations of polysemous words (Kilgarriff 2001). In essence, a language corpus provides the best opportunity to analyze the polysemous words with close reference to the contexts of their use in the language (Cuyckens and Zawada 2001). They help us to specify all contextual frames as well as identify contextual senses. The information obtained from such studies empowers us to verify how the range of meanings obtained from a corpus can match with the existing range of meanings provided in a dictionary.

The *generative lexicon* aims at assigning a structure to words as well as designing a schema that can determine how different senses of a word are combined in specific contexts (Pustejovsky 1995). The success of the schema depends heavily on the fruitful utilization of information derived from a corpus because without reference to a corpus, it fails to account for the varied range of metaphoric meanings a word is able to denote. Moreover, without reference to actual use in a corpus, it cannot make clear-cut distinctions between metonymies and metaphors. Since word meanings are not marked with information about their metaphorical or metonymical sense, the question about how to distinguish literal meaning from non-literal meaning becomes a crucial issue and to answer this question it has to depend on extensive analysis of such words occurring in a corpus.

The figurative use of words is pervasive in all kinds of language use including both informative and imaginative texts. It asks for a considerable investigation into a wide range of fields including general linguistics, psychology, artificial intelligence, and philosophy to understand how figurative senses are triggered from varied lexical frames. Even a few years ago, the majority of studies in this area were guided by the linguistic intuition of investigators without reference to the real use of words. The introduction of the corpus has changed this scenario. Now scholars investigate the figurative use of words in a corpus because a corpus provides the following advantages:

(1) It provides examples of literalness, metaphor, metonymy, polysemy, context-sensitive meaning, etc., to explore relations of figurative usages.
(2) It supplies necessary information to understand inter-annotator agreement on which figurative uses, metaphors, and metonymies are constituted.
(3) It supplies specific linguistic cues for figurative usages, including studies on their frequency, reliability, and evaluation.
(4) It provides information to trace the effect of the domain, genre, or text on the figurative use of words.
(5) It supplies information to design computational models to analyze and interpret figurative uses of words in a language.
(6) It supplies sufficient data for designing cognitive model to process figurative usages of words in a text.

The problem of sense disambiguation of words is one of the central concerns in lexical semantics, language technology, and language processing (Schütze 1998). It has become increasingly apparent in the approaches adopted in the development of WordNet (Dash et al. 2017). The method used in WordNet is based on the approach

that lists up different senses of a word within a web of 'conceptual interface.' However, it says nothing about how and why these senses are interrelated to each other (Miller et al. 1990). This leads to considerable problems since we fail to comprehend how all these senses of words are conceptually related to each other and how we can understand the sense of a word with reference to the meaning of another word. The problem becomes acuter when the novel uses of words occur quite frequently and when new figurative senses are triggered and tagged with existing senses of the words. Since the figurative use of words is pervasive in normal discourse, source meaning of a word used figuratively is often far removed from the intended or target meaning. A possible way to overcome this problem is to list up all the senses of a word from a corpus as well as design a mechanism that can capture new senses from the existing ones after it identifies the inherent relation of the senses.

13.4 Corpus and Sociolinguistics

The sociolinguistics, an empirical branch of linguistics, depends on language data and information procured from various domains of language use interfaced with diverse social interactions. A majority of studies in this branch of linguistics have been more or less concerned with lexical resources to find answers to some simple one-dimensional queries, such as the interface underlying between language and gender, language and ethnic group, language, and geographical region. So far, the majority of the studies of sociolinguistics have used the limited amount of research-specific language data, and in most cases, these data are hardly put to any kind of systematic sampling or quantitative verification. Even, sometimes, the entire sets of data are detached from their natural background of occurrence. Because of this approach, the observation made in many sociolinguistic studies is considered either skewed or non-realistic.

To overcome such deficiency, modern sociolinguists are using well-formed language corpora as these corpora are providing them with a large amount of naturalistic data, which are used for systematic sampling, quantitative measurement, and empirical verification. Some sociolinguists are also using corpora annotated with various kinds of sociolinguistic information and demographic variables (e.g., *age, gender*, *ethnicity, profession, education, caste, social status, ethnicity*, *time, region, motive*) because such corpora are more useful in varied sociolinguistic studies. For instance, the annotated *Brown Corpus* and the *LOB Corpus* are analyzed to trace the feature of 'masculine bias' in the American English and the British English. From the studies, it has been found that the frequency of use of 'female items' is much lower than the 'male items' in both the corpora. Interestingly, however, the use of the 'female items' is more frequent in the British English than in the American English (Kjellmer 1986). Such studies prove the fact that although the American and the British societies claim to extend equal social status to both male and female members in the countries, their actual patterns of language use reveal some truths that directly contradict their claims.

On the other hand, the study relating to the *Corpus of London Teenagers* (COLT) revealed some new insights about the nature of language use by the London teenagers. The study investigated the nature and form of verbal disputes among the British teenagers in several formal and informal situations. The findings have revealed that female teenagers, when they talk to the members of the same group, are equally strong and agile in using slang, sexual terms, and swearing words similar to male teenagers. However, when they are engaged in informal verbal interactions with the male members and the seniors at home or school, they deploy a finer shade of decency and sobriety in their linguistic interactions (Stenström and Hasund 1996).

Another important area of study in sociolinguistics has been the query about why people try to explain things to others both in speech and writing and how do they do it with and without language. In fact, these issues lead some linguists to explore the nature of complex language games such as mediation, negotiation, dialogue, and conversation. It is assumed that explanation or attribution is one of the most important aspects of human speech since it reveals the ways people normally interact with the environment they live in. To verify such arguments, we need access to speech corpora, because such texts are not possible to reproduce in artificial laboratory situations. Also, we require large corpora because we need to quantify and validate our observation. Data extracted from texts where language occurs quite naturally, such as newspapers, personal diaries, negotiations, company reports, dialogues, classroom talks, police investigations, question-answering, market talks, is the most suitable for identifying the factors that operate for providing explanations in speech and writing (Antaki and Naji 1987).

Language corpora are used to describe language varieties as well as for comparative studies between the varieties. In general, a language variety is compared with the other to understand how they vary across text types, domains, times, regions, age, sex, group, profession, ethnicity, etc. The 'variants' may be procured from different parts of the same corpus (e.g., science fiction texts vs. romantic fiction texts) or from similar parts of different corpora (e.g., science fiction texts of a Hindi corpus vs. science fiction texts of a Bangla corpus). For instance, the *LOB Corpus* contains text samples of the same genres and size of the *Brown Corpus*, and both are sampled with texts produced in the same year. Due to compositional similarity, both the corpora have been used to produce frequency lists of words comparable between the American English and the British English in written form to study linguistic and sociolinguistic issues and aspects of the languages. Since the TDIL corpus of Indian languages follows the same sampling procedure to maximize the degree of comparability, it is a good resource for studying various sociolinguistic issues in the languages of the country.

The availability of comparable corpora makes it possible to compare the use of language in different speech communities. Such corpora are analyzed to determine the cultural differences of respective language users (Lovejoy 1995). For instance, after the compilation of the *LOB Corpus* of the British English, it is used to compare its vocabulary with that of the *Brown Corpus* of the American English. The study has revealed many interesting differences, which goes beyond pure linguistic issues such as spelling, morphology, or words (Leech and Fallon 1992). The study has revealed

many interesting differences in culture of the two speech communities, which are never observed before. For example, the number of words relating to tour and travel is far more frequent in the American English than in the British English, which hints toward the larger size of America and the tendency of the American people to travel across regions in the country and abroad, which is not a marked feature of the British people. Similarly, the frequent use of words and terms relating to crime, murder, and military in the American English than in the British English indicates to the American 'gun culture,' which is not a criterion of the British culture. Such findings seem to suggest that the American culture is much more 'macho' and 'dynamic' than the British culture.

In a similar fashion, a comparative study between the *LOB Corpus*, the *Brown Corpus*, and the *Kolhapur Corpus of Indian English* has revealed some interesting observation to trace the differences in the culture of the respective language communities (Shastri 1988). These corpora are also used to study the structure of sentences (Leitner 1991) as well as the patterns of word combination that differ according to the native and non-native speakers (Cock 1998). The results of such studies established the notion of **Common Core Hypothesis** (CCH), the basic argument of which is that all varieties of English used across countries, do possess certain central fundamental properties in common, which may differ quantitatively rather than qualitatively (Quirk et al. 1985: 142). The availability of more comparable corpora opens up many new and promising areas of sociolinguistic studies to probe into the intricate texture of life of people fabricated with language. The *International Corpus of English* (ICE), which is made with different varieties of English texts used in different countries and which is developed following the same designing principles, may be used as a resource for fruitful research in this area (Nelson et al. 2002).

With the availability of bilingual and multilingual parallel corpora of the same frame, size, text, and composition, the possibility for cross-linguistic and cross-cultural studies is gradually opening up. In a multilingual and multicultural country like India, the availability of such corpora can strengthen inter-regional relations, strengthen national integrity, and enhance brotherhood. For instance, the analysis of trilingual corpora made with Assamese, Bangla, and Odia texts can show how words are mostly derived from the same source, how sentences are similar in construction, how lexical meanings bear the same conceptual similarities, how grammatical properties are similar in function, how languages are mutually comprehensible, and how the language users maintain similar tastes, habits, faiths, and lifestyles, etc. Comparative studies of this kind can help us establish interlingual communication, exchange linguistic and extralinguistic information, and strengthen linguistic bonds among the speech communities.

Till date, only a few corpora are used to carry out research within the area of discourse. The primary objectives of such studies are related to understanding how conversations work with respect to lexical items, idioms, and phrasal units, which perform many relevant conversational functions within specific contexts (Stenström 1994). A conversational text corpus annotated with various kinds of geographical, demographic, and social information can provide better scopes to extend such studies both in discourse and pragmatics (Andersen 1997). It is, however, difficult to find

a corpus of this type primarily because it relies heavily on the actual *contexts* of discourse, which are hardly available in a corpus. The samples of written and spoken text stored in a corpus are actually removed from the actual social and textual contexts. Although we may encode sociolinguistic (e.g., gender, class, region, profession, education, culture, ethnicity, domicile) and discourse information (e.g., events, time, participants, contexts, situations, backgrounds) within a corpus, it is not always easy to infer actual contextual information from an annotated corpus (Graf 1996). Let us hope that more research will be carried out in this direction so that we can find more corpora compiled and annotated with relevant information so that corpus-based research may be carried out in this area.

13.5 Corpus and Psycholinguistics

Language corpora are used as a source of data in some recent research works of psycholinguistics. Researchers have obtained texts and data for laboratory experiments. It has been noted that frequency information of words used in a corpus may be considered rudimentary within several psychological processes of language recognition and understanding. In most cases, a systematically sampled corpus is considered useful for providing reliable information about the frequency of use of words with close reference to their occurrence in different senses as well as in different lexical classes. After observing the usage patterns of words manifested in a corpus, psycholinguists are better positioned to formulate hypotheses about the possible patterns of distribution of lexicon in the human mind.

The most valuable role of a corpus in the psycholinguistic experiment is recognized in the examination of occurrence of errors in natural conversation and dialogic interaction. For instance, the *London-Lund Speech Corpus* is most exhaustively used to study speech errors in natural conversations in the British English. The corpus is able to provide exactly the kind of language data required for such a study. After the texts of the corpus are categorized in a systematic order, frequency of various speech errors are counted and classified to retrieve real estimate on the general frequency of errors in relation to overall linguistic outputs of the speakers (Garnham et al. 1981). Before this study, there was no estimate of the frequency of errors in everyday speech, because such analysis required an adequate amount of data from natural conversations, which was not available. Previous works on speech errors were mostly based on small-scale ad hoc collection of a few spoken text samples obtained through interviews with the informants.

The use of corpus is also approved in language pathology—an important domain of psycholinguistic research (MacWhinney 1991). The primary aim of this area of research is to understand why and how people suffer from various linguistic deficiencies. Generally, such studies require a large amount of language data produced by linguistically impaired people while interacting in various situations. Till date, most of the studies on language data produced by linguistically impaired people lacked the feature of quantified representation for reliable observation. The generation of

special speech corpus, however, has partially fulfilled the need for accurate 'abnormal speech data,' which helps us develop methods for testing error patterns and identifying factors that create problems in the cognitive processing of a language. Although a few works have been done so far with the specialized corpus, their potential importance in linguistic pathological research cannot be ignored. For instance, the *CHILDES* database that contains a large amount of data produced by linguistically impaired and normal children has been processed and analyzed to understand the underlying problems in these areas (Biber et al. 1998: 177).

13.6 Corpus and Stylistics

The availability of corpus made with texts belonging to different genres, domains, authors, media, etc., has opened up many new areas for research into stylistics (Stubbs 1996). Within a broader canvas of stylistics, researchers are interested in individual text types as well as texts composed by authors with specific stylistic criteria. For instance, scholars are interested to find out the basic stylistic differences reflected in the texts composed by the writers of one country with that of the other countries (Wilson 1992). Similarly, scholars are interested to know how the writings of one generation of writers or a group of writers are stylistically different from the writings of other generation of writers or the other group of writers. Thus, diversified and comparative studies in stylistics are possible when the scholars have access to large synchronic and diachronic corpora representing texts marked various stylistic features they consider relevant for their studies (Eskénazi 1993; Miller 2001; Biber 1988).

Although researchers are interested to investigate broader issues such as genre and type of texts, quite often they are also willing to deal with stylistic factors in a language to concentrate on some specific features of certain text types (Biber 1986). They are interested to know how and to what proportion the language used in scientific texts varies stylistically from the language used in media texts. Such investigations require large corpus for faithful analysis and verifiable inferences (Biber 2002). A general corpus, as well as special corpus, becomes an important source of data because both of them can serve as a frame of reference to make comparisons within them as well as with other types of the corpus (Halliday 1987, 1989).

Any object-oriented study in stylistics also requires statistically verified information to back up the judgments that appear subjective rather than objective to the investigators (Hoffman 1955). The analysis of corpus is effective for tracking changes in writing styles, identifying the patterns of word selection, tracing the patterns of text narration and topic description, specifying the patterns of sentence formation and content representation, etc. For authorship attribution as well as for defining an author's particular style of writing, we can use corpus made with writings of that particular author to identify how the author leans toward different ways of putting things (e.g., personal vs. general, technical vs. non-technical, formal vs. informal). It requires comparisons to be made not only internally within the author's own work

but also with the works of other authors or the norms of the language or variety as a whole (deHaan 1997). This exhibits not only the style of writing of the author under consideration but also the style in which the text is composed. Such stylistic investigation needs various statistical data and examples best available from a corpus.

The comparative stylistic analysis of texts composed by a single author can show how he shifts across techniques, narration, vocabulary, sentence, style, etc., while dealing with different types of content (Elliott and Valenza 1996). For example, if we make a comparative study of the different types of prose text produced by Rabindranath Tagore (e.g., short stories, novels, essays, travelogues, personal letters), we can find that his modes of narration, techniques of sentence construction, manners of text representation, choices of vocabulary, etc., vary from text to text. Such a study can show how a single author varies in the application of different stylistic nuances based on the topic or type of a text.

The TDIL text corpus available in the Indian languages is full of information on various genres, time frames, and text types (Dash 2007). A sensible use of texts of this corpus can open up many new possibilities of research in stylistics within and across the Indian languages. At the initial stage, we may use these texts for simple comparisons about the variety of narrative styles observed within single sample text of a language. In future, it may be extended to several other text types within or across the languages. In essence, due to easy comparability, the TDIL corpus is a valuable resource for studying different linguistic features and styles within and across multilingual frames. This corpus may also be used to identify features of individual text types as well as to attribute authorship of texts to particular authors.

13.7 Conclusion

Linguistics, because of its never-ending magic, has always been treated as an enigmatic field of study that invites people to probe into various dimensions of human life and society through the language they use. The long history of the linguistic study of the last three millenniums has established the fact that linguistics will continue to thrive with life and will serve as long as the people use language as the most powerful tool for expressing their minds and establishing communication with others.

For centuries, linguistics has been an area of utmost curiosity, attention, and investigation among the philosophers, lexicologists, grammarians, rhetoricians, and others. It has been studied from various angles with diverse goals and missions. Starting from the introspective analysis, it has been a subject of discussion of various kinds within the descriptive analysis, normative proposition, comparative method, generative framework, and intuitive reflection (Chomsky 1957). In every field, several rules and methods are formed, instructions and prescriptions are proposed, and principles and theories are generated. All these have generated renewed interest among the people about language in every phase of human civilization, and attempts have been made to explore more and more into the language.

Corpus linguistics is one such method. It is an area that focuses on 'language in use' with reference to users and contexts. It shows with evidence how people use language to mark unique identity, show solidarity, express ideas, share information, cultivate knowledge, promote culture, preserve history, build solidarity, nourish communities, and draw differences. Such activities are going on generations after generations. A corpus of its wide spectrum invites us to explore all these issues and aspects of language reflected in various fields of linguistic interaction. The underlying truth of a corpus is that evidence of language use carries higher value and importance as it directly focuses on 'authenticity' of a language that tends to change over times.

The present trend of linguistic research all over the world is tilted toward empirical model pillared on language corpus. Even generative linguists are turning their attention toward corpus to verify existing theories and principles proposed in the generative frame. Thus, the combination of computer and corpus has brought in a deluge into the life of the so-called dull and lifeless fields of linguistics. Perhaps, it will emerge as the most promising area in future for the survival and growth of linguistics as a discipline.

References

Alanko, P.K. 2000. Mechanisms of Semantic Change in Nouns of Cognition: a General Model. In *Lexicology, Semantics, and Lexicography*, ed. J. Coleman and C.J. Kay, 29–52. John Benjamins: Amsterdam-Philadelphia.

Andersen, G. 1997. They Like Wanna See 'like' How We Talk and All That: The Use of 'like' as a Discourse Marker in London Teenage Speech. In *Corpus-based Studies in English*, ed. M. Ljung, 37–48. Rodopi: Amsterdam-Atlanta, GA.

Antaki, C., and S. Naji. 1987. Events Explained in Conversational 'Because' Statements. *British Journal of Social Psychology* 26: 119–126.

Biber, D. 1986. Spoken and Written Textual Dimensions in English. *Language* 62 (4): 384–414.

Biber, D. 1988. *Variation Across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, D. 2002. Speaking and Writing in the University: A Multidimensional Comparison. *TESOL* 36 (1): 9–48.

Biber, D., S. Conrad, and R. Reppen. 1998. *Corpus linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Bouillon, P., and F. Busa (eds.). 2001. *The Language of Word Meaning*. Cambridge: Cambridge University Press.

Cabanillas, I.C., and C.T. Martínez. 2002. The Horse Family: On the Evolution of the Field and Its Metaphorisation Process. In *A Changing World of Words: Studies in English Historical Lexicography, Lexicology and Semantics*, ed. D.E.J. Vera, 229–254. Amsterdam: Rodopi.

Chomsky, A.N. 1957. Review of *Verbal Behavior* by B.F. Skinner. *Language* 29 (1): 26–58.

Cock, S.D. 1998. A Recurrent Word Combination Approach to the Study of Formulae in the Speech of Native and Non-native Speakers of English. *International Journal of Corpus Linguistics* 3 (1): 59–80.

Coleman, J., and C.J. Kay (eds.). 2000. *Lexicology, Semantics, and Lexicography: Selected Papers from the 4th G.L. Brook Symposium*. Amsterdam: John Benjamins.

Cowie, C., and C.D. Puffer. 2002. Diachronic Word-formation and Studying Changes in Productivity Over Time: Theoretical and Methodological Considerations. In *A Changing World of Words:*

Studies in English Historical Lexicography, Lexicology and Semantics, ed. D.E.J. Vera, 410–437. Amsterdam: Rodopi.

Cuyckens, H., and B. Zawad (eds.). 2001. *Polysemy in Cognitive Linguistics*. Amsterdam: John Benjamins.

Dash, N.S. 2007. Indian Scenario in language Corpus Generation. In *Rainbow of Linguistics: Vol. 1*, ed. N.S. Dash, P. Dasgupta, and P. Sarkar, 129–162. Kolkata: Media Publication.

Dash, N.S., P. Bhattacharyya, and J.D. Pawar (eds.). 2017. *The WordNet in Indian Languages*. Singapore: Springer.

deHaan, P. 1997. Some Experiments in Authorship ATTRIBUTION. In *From Aelfric to the New York Times*, ed. U. Fries, V. Mü, and P. Schneider, 125–137. Amsterdam: Rodopi.

Elliott, W., and R. Valenza. 1996. And Then There Were None: Winnowing The Shakespeare Claimants. *Computers and the Humanities* 30 (3): 1–56.

Eskénazi, M. 1993. Trends in Speaking Styles Research. In *Proceedings of the 3rd European Conference on Speech Communication and Technology*, vol. 1, 501–512. Eurospeech'93, Berlin, Germany.

Fillmore, C.J., and B.T.S. Atkins. 2000. Describing Polysemy: The Case of 'Crawl'. In *Polysemy*, ed. Y. Ravin and C. Leacock, 91–110. New York: Oxford University Press Inc.

Francis, W.N. 1982. Problems of Assembling and Computerizing Large Corpora. In *Computer Corpora in English Language Research*, ed. S. Johansson, 7–24. Norwegian Computing Centre for the Humanities: Bergen.

Garnham, A., R. Shillock, G. Brown, A. Mill, and A. Cutler. 1981. Slips of the Tongue in the London-Lund Corpus of Spontaneous Conversation. *Linguistics* 19: 805–817.

Gevaert, C. 2002. The Evolution of the Lexical and Conceptual Field of ANGER in Old and Middle English. In *A Changing World of Words: Studies in English Historical Lexicography, Lexicology and Semantics*, ed. D.E.J. Vera, 275–299. Amsterdam: Rodopi.

Graf, D. 1996. *Relative Clauses in Their Discourse Context: A Corpus-Based Study*. Unpublished M.A. Thesis: Freiburg.

Halliday, M.A.K. 1987. *Spoken and Written Modes of Meaning, Comprehending Oral and Written Language*. San Diego, CA: Academic Press.

Halliday, M.A.K. 1989. *Spoken and Written Language*. Oxford: Oxford University Press.

Hoffman, C. 1955. *The Man Who was Shakespeare*. New York: Julias Messner Inc.

Hofland, K., and S. Johansson. 1982. *Word Frequencies in British and American English*. Bergen: Norway Computing Centre for the Humanities.

Hundt, M. 1997. Has British English Been Catching up with American English Over the Past Thirty Years? In *Corpus-Based Studies in English: Papers from the 17 International Conference on English-Language Research Based on Computerised Corpora*, ed. M. Ljung, 129–151. Amsterdam: Rodopi.

Kilgarriff, A. 2001. Generative Lexicon Meets Corpus Data: The Case of non-Standard Word Uses. In *The Language of Word Meaning*, ed. P. Bouillon and F. Busa, 312–328. Cambridge: Cambridge University Press.

Kjellmer, G. 1986. 'The lesser man': Observations on the Role of Women in Modern English Writings. In *Corpus Linguistics II*, ed. J. Aarts and W. Meijs, 163–176. Amsterdam: Rodopi.

Leech, G., and R. Fallon. 1992. Computer Corpora: What Do They Tell US About Culture. *International Computer Archive of Modern English Journal* 16: 29–50.

Leech, G., B. Francis, and X. Xu. 1994. The Use of Computer Corpora in the Textual Demonstrability of Gradiance in Linguistic Categories. In *Continuity in Linguistic Semantics*, ed. C. Fuchs and B. Vitorri, 31–47. John Benjamins: Amsterdam and Philadelphia.

Leitner, G. 1991. The Kolhapur Corpus of Indian English: Intravarietal Description and/or Intervarietal Comparison. In *English Computer Corpora: Selected Papers and Research Guide*, ed. S. Johansson and A.-B. Stenström, 215–232. Berlin: Mouton de Gruyter.

Lewis, D.M. 2002. Rhetorical Factors in Lexical-Semantic Change: The Case of 'At Least'. In *A Changing World of Words: Studies in English Historical Lexicography, Lexicology and Semantics*, ed. D.E.J. Vera, 525–538. Amsterdam: Rodopi.

Lovejoy, J. 1995. Prepositions in British and American English—A Computer-Aided Corpus Study. *Arbeiten aus Anglistik und Amerikanistik* 20: 55–74.
MacWhinney, B. 1991. *The CHILDES Project: Tools for Analyzing Talk*. Hillsdale, N.J.: Lawrence Erlbaum.
Miller, J. 2001. Spoken and Written Language. *Pragmatic Organisation of Discourse in the Languages of Europe*, G ed, 56–67. Berlin: Mouton de Gruyter.
Miller, G., Beckwith, R., Fellbaum, C., Gross, D., and K. Miller. 1990. *Five Papers on WordNet.* Cognitive Science Laboratory, Princeton University, Princeton.
Mindt, D. 1991. Syntactic Evidence for Semantic Distinctions in English. In *English Corpus Linguistics: Studies in Honour of Jan Svartvik*, ed. K. Aijmer and B. Altenberg, 182–196. London: Longman.
Nelson, G., S. Wallis, and B. Aarts. 2002. *Exploring Natural Language: Working with the British Component of the International Corpus of English*. Amsterdam: John Benjamins.
Poch, A., and I.V. Clavera. 2002. The Rise of New Meanings: A Historical Journey Through English Ways of 'Looking At'. In *A Changing World of Words: Studies in English Historical Lexicography, Lexicology and Semantics*, ed. D.E.J. Vera, 563–571. Amsterdam: Rodopi.
Pustejovsky, J. 1995. *The Generative Lexicon*. Cambridge, Mass.: MIT Press.
Quirk, R., S. Greenbaum, G. Leech, and J. Svartvik. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.
Ravin, Y., and C. Leacock (eds.). 2000. *Polysemy: Theoretical and Computational Approaches*. New York: Oxford University Press Inc.
Schütze, H. 1997. *Ambiguity Resolution in Language Learning: Computational and Cognitive Models*. Cambridge: Cambridge University Press.
Schütze, H. 1998. Automatic Word Sense Disambiguation. *Computational Linguistics* 24 (1): 97–123.
Shastri, S.V. 1988. The Kolhapur Corpus of Indian English and Work Done on Its Basis So Far. *International Computer Archive of Modern English Journal* 2: 15–26.
Stenström, A.-B. 1994. *An Introduction to Spoken Interaction*. London: Longman.
Stenström, A.B., and I.K. Hasund. 1996. Girls' conflict talk: a sociolinguistic investigation of variation in the verbal disputes of adolescent females. In *A Study from COLT Corpus of London teenager language*. The University of Bergen. Paper presented at ICAME, Stockholm.
Stubbs, M. 1996. *Text and Corpus Analysis: Computer-Assisted Studies of Language and Culture*. Oxford: Blackwell.
Teubert, W. 2000. Corpus linguistics—A Partisan View. *International Journal of Corpus Linguistics* 4 (1): 1–16.
Tissari, H. 2000. Five Hundred Years of LOVE: A Prototype Semantic Analysis. In *Lexicology, Semantics, and Lexicography*, ed. J. Coleman and C.J. Kay, 127–156. Amsterdam: John Benjamins.
Vera, D.E.J. (ed.). 2002. *A Changing World of Words: Studies in English Historical Lexicography, Lexicology and Semantics*. Amsterdam: Rodopi.
Wilson, A. 1992. The Usage of 'Since': A Quantitative Comparison of Augustan, Modern British, and Modern Indian English. *Lancaster Papers in Linguistics* 80: 17–32.

Web Links

https://www.uploady.com/#!/download/Oz2gb165H7N/3TU0alSAhrH2J4SG.
http://www.euralex.org/elx_proceedings/Euralex2004/.
http://www.cs.mu.oz.au/research/lt/nlp06/materials/Baldwin/intro.pdf.
http://semarch.linguistics.fas.nyu.edu/barker/Research/barker-lexical.pdf.
http://web.stanford.edu/class/linguist1/Rdgs/JM19.pdf.

http://dingo.sbs.arizona.edu/~hharley/PDFs/WordsBook/Chapter6.pdf.
http://web.stanford.edu/class/linguist1/Slides/hofmeister-slides.pdf.
http://people.umass.edu/cec/languagecomprehension.pdf.pdf.
http://www.mpi.nl/world/materials/publications/levelt/Levelt_Psycholinguistics_1992.pdf.
http://www.mantex.co.uk/2009/09/13/english-language-stylistic-analysis/.
http://digitalhumanities.org/companion/view.
http://teach-grammar.com/wp-content/uploads/2012/07/2010+-Grammar-and-Stylistics.pdf.

Chapter 14
Language Corpora: The Indian Scenario

Abstract The humble goal of this chapter is to refer to some of the achievements in the area of Indian languages corpora generation and lexical databases compilation, which have been done for a few Indian languages within last two and half decades. We shall also try to refer in this chapter some works which are still in the process of continuation for the development of corpora and lexical database for the Indian languages. What is most satisfying is the involvement and active participation of a large number of renowned institutions and individuals of the country in such works due to which these have drawn a considerable amount of attention and approval across the globe. In essence, in this chapter, we shall make a short survey on the development of monolingual corpora and parallel translation corpora, which are developed through some individual attempts or joint enterprise across the country.

Keywords Indian languages corpora · Kolhapur corpus of Indian English · TDIL corpus · Indian languages corpora initiative · LDC-IL · Monolingual corpora Parallel corpora · Language archive

14.1 Introduction

Since the history of digital corpora and lexical database generation in the Indian languages is punctuated with conceited objections, discrete oppositions, and multilingual diversities, this sketchy survey is, therefore, a short chronicle about the major milestones achieved so far with a welcome invitation to those recent works that are also making honest efforts to contribute toward the development of tools, technology, and resources for the Indian languages. The motivation that inspires the new generation of Indian scholars to work in collective enterprise is also appreciated here, as divergent schemes of work undertaken in the domain of corpus and lexical database generation will eventually turn India into a 'resource-rich' country—a collective dream cherished for long by the entire NLP community of the country. The functional adequacy of these resources and language technology tools is, however, an

N. S. Dash and L. Ramamoorthy, *Utility and Application of Language Corpora*,
https://doi.org/10.1007/978-981-13-1801-6_14

important criterion which is yet to be achieved within a pan-Indian frame by which the entire nation can benefit.

During last three decades, it has been one of our national goals to work for language processing and language technology for the Indian languages, as we lack in properly developed, digitized, and processed language corpora based on which advanced user-friendly tools, systems, and techniques can be developed for general access across all sociocultural boundaries and register variables of the country (Dash et al. 2016). Keeping this goal in view, in this chapter, we shall make an effort to refer to those major resources that are already developed in the Indian languages and that contribute further toward the growth of NLP research activities in Indian languages.

The chapter is organized in the following manner. In Sect. 14.2, we refer to the Kolhapur Corpus of Indian English (KCIE), the first digital text corpus in the Indian language; in Sect. 14.3, we discuss the TDIL (Technology Development in Indian Languages) corpus, the first pan-Indian general corpus in the Indian languages; in Sect. 14.4, we discuss the ILCI corpus (Indian Languages Corpora Initiative), the first parallel translation corpus in twelve Indian languages; in Sect. 14.5, we refer to the LDC-IL, the first digital archive of the Indian languages—a national linguistic depository that collects, stores, manages, and disseminates linguistic resources, tools, and systems for national causes; and finally, in Sect. 14.6, we refer to some private and public agencies where corpus generation works in the Indian languages are carried out.

14.2 KCIE: Kolhapur Corpus of Indian English

The development of indigenous corpus in the Indian languages began, in the true sense of the term, in 1991 when a concerted effort was initiated under the aegis of the *Department of Electronics and Technology* (DeitY), *Government of India* to generate digital text corpora for most of the Indian languages for various NLP activities. This effort, however, should not be considered first of its kind in India, since the work of corpus generation in any Indian language began some time ago at the individual level at the *Shivaji University, Kolhapur*, *India* (Shastri 1988). To the best of our knowledge, the *Kolhapur Corpus of Indian English* (KCIE) is the first Indian corpus, which was systematically developed following the norms adopted for the development of the *Brown Corpus* (Francis and Kucera 1964) and the *LOB (Lancaster-Oslo-Bergen) Corpus* Atwell et al. (1984).

The KCIE, as the name indicates, was made with written text samples of modern *Indian English* with a goal of making a cross-comparison between the *British English*, the *American* English, and the *Indian English* (Shastri 1988). At present, it is available at the *International Computer Archive of Modern English* (ICAME), while details of the corpus are available at the *University of Bergen, Norway*. It consisted of approximately one million words of *Indian English* drawn proportionally from the text materials published in the year 1978. The text samples are collected from 15 different text categories to make it maximally comparable to the *Brown Corpus* and

the *LOB Corpus*. The text samples are manually inserted into computer following the *American Standard Code for Information Interchange* (ASCII) to make the corpus data maximally retrievable and accessible by the end users. At present, the *KCIE* is included as a representative sample of the *Indian English* in the *International Corpus of English* (ICE) and used as an authentic linguistic resource for studies on *Indian English*.

Although the *KCIE* was initially planned to make it maximally comparable to the *Brown Corpus* and the *LOB Corpus*, there are, indeed, some deviations in the KCIE from these two corpora due to some practical problems. The first deviation was that it failed to match the *Brown Corpus* and the *LOB Corpus* with respect to synchronicity. While the *Brown Corpus* and the *LOB Corpus* stored sample of English texts published in the year 1961, the *KCIE* stored samples of English texts published in the year 1978. In spite of differences in years, the designers of the *KCIE* were successful in preserving maximum comparability among the three corpora and maintaining parallelism in the representation of samples across genres and text types.

However, the most notable difference of the *KCIE* from the other two corpora was that the required balance in text representation was titled toward 'informative prose.' This seemed inevitable as the texts available in the area of 'imaginative prose' fell short of the number required for maintaining the balance in the composition. Also, importance on texts of short stories as against that of full-length novels tilted the balance toward informative texts. Such skewed representation of texts is hardly noted in case of other text categories, as the table shows (Table 14.1).

Since the *KCIE* was meant to represent the *Indian English* used in printed and published documents, the representative text samples were accumulated in a simple, stratified, and random manner from the following sources:

(a) **Printed books**: 140 sample texts from 1200 titles of different disciplines,
(b) **Government documents**: 37 sample texts from central and state government documents,
(c) **Press materials**: 53 sample texts from 6 national and 15 regional English newspapers,
(d) **Periodicals**: 170 samples texts from almost all fields and disciplines.

The referential value of the *KCIE* has been attested in its faithful representation of sample texts for describing the actual form and nature of the *Indian English*. The referential value of this corpus may be understood from the contribution it has made in the study of Indian English.

(a) It has succeeded to exhibit distinct texture of the *Indian English* enriched with unique lists of Indian words and terms not found in other two corpora (Wilson 1992).
(b) It has faithfully reflected on the differences in syntactic patterns (Shastri 1992, 1996) and semantic loads of the *Indian English* (Marco 2006).
(c) It has exhibited the independence of the *Indian English* from the overwhelming shadow of the *British English* (Mukherjee 2002) or the *American English*.
(d) It has become successful to exhibit the 'Indianness' of the *Indian English* as a notable phenomenon of post-Independent India that has achieved a distinct and discernible linguistic identity within last few decades (Leitner 1991).

Table 14.1 Composition of the *Brown*, *LOB,* and *KCIE*

Text Categories		Texts in each category		
		Brown	LOB	KCIE
Informative Text	Press: reportage	44	44	44
	Press: editorial	27	27	27
	Press: reviews	17	17	17
	Religion	17	17	17
	Skill, trade, and hobbies	36	38	38
	Popular lore	48	44	44
	Belles letters	75	77	70
	Miscellaneous (foundation reports, industry reports, government documents, college catalogue, industry origin)	30	30	37
	Learned scientific writings	80	80	80
Imaginative Texts	General fiction	29	29	58
	Mystery and detective fiction	24	24	24
	Science fiction	6	6	2
	Adventure (western fiction)	29	29	15
	Romance and love story	29	29	18
	Humor	9	9	9
	Total	500	500	500

Even then, it is clear that the *KCIE* cannot claim to be maximally representative of the *Indian English* since it is built to ensure maximum comparability with other English corpora (i.e., *Brown* and *LOB*) rather than developing a maximum representative text database of the *Indian English*. In spite of several limitations and shortcomings, the *KCIE* has registered immense referential value in Indian linguistics in general, because it had provided a great opportunity to analyze the form and features of the *Indian English* with close reference to a real-life use of the language. In fact, the availability of the *KCIE* has made it possible to observe the distinct features of the *Indian English*, which has not been attested before so elaborately (Leitner 1991). Till date, it is used as one of the primary resources of the *Indian English*, and it is made open to researchers for various linguistic studies and investigations.

14.3 The TDIL Corpus

The TDIL (*Technology Development for the Indian Languages*) scheme, a dear daughter of the *Ministry of Communication and Information Technology* (MCIT), *Government of India*, was conceived in the year 1991 to provide technology solutions

for the Indian languages. The primary objectives of this program were the followings (Dash 2007):

(a) To develop information processing techniques and tools for all Indian languages,
(b) To facilitate human–machine interactions crossing all language barriers,
(c) To create, share, and access multilingual knowledge resources,
(d) To integrate resources and tools for devising innovative user-friendly products and services.

The work of generating text corpora in the Indian languages began in the year 1992 under the guidance of the TDIL team. The initial goal was to construct large text corpora in digital form for all the major Indian languages included in the *8th Schedule* of *The Constitution of India*. The mission was to collect representative text samples of three million words from each of the Indian languages considered in the project by way of adopting a well-defined uniform sampling method through which equal amount of data from each of the languages was to be collected from different disciplines and subject areas (Murthy and Despande 1998).

The project was collectively shouldered by six premiere institutes and universities across the country. Each institute/university was entrusted with the task of collecting text corpora from three to four languages assigned to each organization as well as process and analyze these corpora for developing NLP tools for spelling checking, morphological analysis, grammar checking, machine translation, digital lexical database generation, and information retrieval, etc. The institutes and universities that took part in this collective work are listed in Table 14.2.

Although not spelt out, several other distant motives of the project were also visualized and these may be reproduced in the following ways (Murthy and Despande 1998).

- To develop representative text corpus for each of the Indian languages that would stand as representative of the language,
- To develop language processing tools, systems, and techniques to facilitate the man–machine interactional interfaces,
- To enhance multilingual knowledge management and sharing across the Indian languages including English,

Table 14.2 Corpus creation in the Indian languages in TDIL project

Languages	Institutes/Universities
English, Hindi, and Punjabi	Indian Institute of Technology, Delhi
Tamil, Malayalam, Telugu, and Kannada	Central Institute of Indian Languages, Mysore
Marathi and Gujarati	Deccan College, Pune
Odia, Bangla, and Assamese	Indian Institute of Applied Language Sciences, Bhubaneswar
Sanskrit	Sampurnananda Sanskrit University, Varanasi
Urdu, Sindhi, and Kashmiri	Aligarh Muslim University, Aligarh

- To promote the development of techniques and systems for cross-lingual investigations, studies, and research,
- To generate linguistic recourses for machine translation, information retrieval, and language education,
- To build machine translation systems among the Indian languages and from English to Indian languages,
- To develop man–machine interactional interfaces needed for information exchange across all the Indian languages,
- To develop computer-assisted language education systems in the Indian languages.

Keeping these immediate and distant objectives in view, the TDIL project was started in the year 1992 for generating machine-readable corpora in the Indian languages. At the initial stage, the corpus developers were slightly skeptical about the practical relevance of digital corpora in context of the Indian languages. It was also necessary to decide whether the proposed digital corpora would contain the samples of written texts or the samples of transcripted spoken texts or both types of text. Although speech has been widely acknowledged as a much better representative of the basic structure and fundamental organization of a language (Biber 1986; Halliday 1989), the corpus developers were forced to focus mainly on written texts due to certain technical constraints that were very much prevalent at the time when the project was initiated. With dubious conviction, the work for corpus generation started keeping the following questions in view:

(a) Why there was a need for developing digital corpora in the Indian languages?
(b) Who were to develop these corpora and how?
(c) How large would these corpora be in a total number of words?
(d) Who would be using these corpora, where, and how?
(e) What kind of text would be included in these corpora?
(f) How much time would be required to develop these corpora?
(g) Which time span would be represented in these corpora?
(h) How would data of actual language use be collected?
(i) What kind of formalities would be adopted for data storage, text representation, and data management?

Many such questions were addressed satisfactorily at the time of actual corpus generation task (Dash 2007). The generation of the TDIL corpus also involved some other issues, such as size of corpora, choice of documents, collection of text materials (e.g., books, newspapers, magazines, periodicals), selection of text samples, sorting process of texts, manner of page selection (e.g., random, regular, selective), determination of end users, manners of data input, methods of corpus sanitation, management of corpus databases, release of corpora for public access, question of copyright of texts. (Atkinset al. 1992). Most of the issues were addressed adequately with reference to the Indian language corpora in the following manners:

(a) **Corpus size**: Each corpus contained three million or more words randomly selected from printed and published texts of each Indian language.

(b) **Text representation**: Text samples were selected from all kinds of texts available during that time.
(c) **Text selection**: Random and selective methods were adopted for text selection.
(d) **Time span**: Texts produced between the years 1981 and 1990 were selected.
(e) **Documents**: Text samples were subject-wise, year-wise, and name-wise.
(f) **Newspapers**: Texts from newspapers were also classified based on names, years, months, and dates.
(g) **Books**:Sample texts were collected from 85 subjects or disciplines.
(h) **Authors**: Texts produced by all authors irrespective of caste, creed, gender, age, ethnicity, influence, race, locality, merit, education, profession, etc., were taken.
(i) **Target users**: People from all walks of life were the target users. They could use the corpora for any language-related study, investigation, or application.

Following these strategies, each group was able to generate a text corpus of three million words in each of the Indian languages included in the project. The project came to an end by the end of March 1995, when each of the Indian languages included in the project was rich with a corpus of modern texts ready for use in all kinds of linguistic works. These corpora are now under the custody of the *Central Institute of Indian Languages* (CIIL), Mysore, which is entitled to the dissemination of the resource to interested Indian scholars and institutes for research and development purposes.

However, the most unfortunate thing is that till date these corpora have not come into much use due to certain technical and legal constraints. The copyright issue is one of them. Moreover, the creation of the corpora in the ISCII (*Indian Standard Code of Information Interchange*) has made these databases nearly obsolete because of their complex conversion compatibility to the Unicode. Although corpora of some of the languages (e.g., Hindi, Bangla, Odia, Marathi, Tamil, Telugu) are converted into Unicode by individual efforts, the rest of the corpora still remains in ISCII, thereby mostly non-utilizable. If these were converted into the Unicode-compatible texts, these would have been invariably used by people for various works of NLP and LT, besides regular works of linguistics. Anybody who is interested in Indian languages and linguistics could have used these corpora because these corpora are developed in such a way that these could easily provide information of the following types:

(a) Present a better scope for studying variations of the relation of linguistic elements across all text types in a language.
(b) Provide a wider spectrum of the language used to study the frequency of occurrence of various linguistic items in the language.
(c) Provide a better opportunity for exploring the behavioral patterns of various language elements in regular language texts.
(d) Confirm the increment in the number of citations to provide facilities for systematic classification of linguistic items in terms of usage and meaning.
(e) Assure better opportunity for obtaining statistical results of the language elements for making various linguistic and extralinguistic observations.

(f) Give a wider spectrum for studying the patterns of use of lexical items to make a generalization about the grammar of a language (Sinclair 1991),
(g) Provide larger scope for a faithful descriptive study on the patterns of use of compounds, lexical collocations, phrases, clauses, sentences, technical and scientific terms, idioms and proverbial expressions, in the language, etc.
(h) Provide opportunities to track the coinage of new words and their patterns if used in different domains of language use.
(i) Give ample scope to track variations of a sense of words triggered by the contexts of occurrences at the both sentence and larger frames.
(j) Give scope to generate a lexical database of different types to the analysis to provide insights into the patterns of formation of words, compounds, idiomatic expressions, and phrases in a language.
(k) Supply authentic analysis and citation of examples of spelling variation—a real critical issue in some of the Indian languages like Bangla, Odia, Tamil.

These are some of the issues, which could have been properly investigated with data and information available in these corpora. In our view, besides many other applications of these corpora in linguistics and language technology, these are perhaps the best resource for investigating the patterns of lexical use in the language—an essential research component in word-sense disambiguation (Hanks 2004) as well as in semantic-type identification (Hanks and Pustejovsky 2005).

By executing the TDIL corpus project, the Indian scholars have definitely achieved some milestones, if not all. Also, they have made some achievements in the form of basic tools, software, and fonts for almost all the major Indian languages. In case of generating language resources, the TDIL has surely made a hallmark by generating corpora in most of the Indian languages (although small in size with regard to a number of words). At present, the ministry is striving hard to run some more projects dedicated toward the development of linguistic tools, systems, and resources of different types to address different needs of the Indian NLP activities. The information elicited from the *Viswabharat* (2011)—an *Indian Technology Newsletter*—shows that the TDIL activities are focusing on the development of the following resources and tools for the Indian languages:

(a) English to Indian languages machine translation,
(b) Indian language to Indian language machine translation,
(c) Sanskrit to Hindi machine translation,
(d) Document analysis and recognition,
(e) Online handwriting recognition,
(f) Cross-lingual information access and retrieval,
(g) Speech corpora and speech-related technologies,
(h) Parallel text corpora in national languages.

If these milestones are achieved, we can surely claim that India has made a giant leap toward the technology development for her languages.

14.4 ILCI: Indian Languages Corpora Initiative

In 2015, a team of scholars across Indian institutes and universities has completed the two phases of the ILCI (*Indian Languages Corpora Initiative*) project where they have developed tagged parallel translation corpora for twenty-three Indian languages, keeping Hindi as the source language and other Indian languages as the target language. In the first phase (2009–2012), the team has generated 50K POS-tagged parallel sentences in each of the twelve Indian languages involved in the project covering two major domains: health and tourism. The total strength of the corpus in each of the languages is 600K annotated sentences with each sentence having an average length of 16 or more words (Jha 2010). Some information about the member institutes and their respective languages (in alphabetical order) involved in the first phase (ILCI-1) is given in Table 14.3.

The most important feature of the ILCI database is that parallelism in text types and sentence structure is preserved at the highest possible level across all the Indian languages—making this corpus database an indispensable resource for cross-lingual information retrieval, Core Grammar development, machine translation, WordNet design, common lexical database generation, and cross-cultural studies and investigation. Another important contribution of the ILCI project is the development of a nationally approved Part-of-speech (POS) tagset known as the BIS (*Bureau of Indian Standard*) Tagset—a benchmark POS standard proposed to be adopted and used for all the Indian languages. An interesting by-product of this project is the generation of several bilingual parallel lexical databases—that will eventually lead to the compilation of domain-specific digital bilingual and multilingual dictionaries in all the Indian languages involved in the project. At present, the ILCI corpus is made available in a Unicode format for general access from the TDIL *Data Centre* of the *DIT, Ministry of the Information and Communication Technology, Government of India*.

Table 14.3 Languages and institutes of ILCI-1 (2009–2012)

Language	Institute/University
Bangla	Indian Statistical Institute, Kolkata
English	Jawaharlal Nehru University, New Delhi
Gujarati	Gujarat University, Ahmadabad
Hindi	Jawaharlal Nehru University, New Delhi
Konkani	Goa University, Goa
Malayalam	IIITM-Kerala, Trivandrum
Marathi	Indian Institute of Technology, Mumbai
Odia	Utkal University, Bhubaneswar
Punjabi	Punjabi University, Patiala
Tamil	Tamil University, Thanjavore
Telugu	Dravidian University, Kuppam
Urdu	Jawaharlal Nehru University, New Delhi

The second phase of the ILCI project (ILCI-2) was more challenging and promising. It started in March 2012 with additional eleven Indian national languages and ended in September 2015. It added 1,700,000 new sentences (1,100,000 sentences from 11 new languages including the four northeast Indian languages and 600,000 sentences from the existing 12 languages). The total strength of the corpus after ILCI-2 is estimated to be approximately 36,800,000 POS-annotated words including all 23 languages in the following four domains: tourism, health, agriculture, and entertainment. This project showed how a mountain can be made out of the sand grains scattered in the dusty plane (Bansalet al. 2013).

14.5 The LDC-IL

The language promotion and maintenance department of the *Ministry of Human Resource Development* (MHRD), *Government of India*, has formed a digital archive for the Indian languages known as the *Linguistic Data Consortium for Indian Languages* (LDC-IL) and has invited the *Central Institute of Indian Languages* (CIIL), Mysore, to host and operate the consortium in a collective manner involving academic institutions and individuals across the country. The consortium, which is set up following the line of the *Linguistic Data Consortium* (LDC) of the *University of Pennsylvania*, USA, is assigned with many important tasks of linguistics and language technology for the Indian languages, such as generation, storage, management, and dissemination of the Indian language corpora; generation and storage of generalized and special lexical databases of the Indian languages; creation of forums for researchers in India and abroad to work on the Indian languages; development of linguistic tools and products; publication of linguistic resources; and building up a large pool of scholars as human resources for all the Indian languages. Besides these long-distant goals, there are some immediate goals also that LDC-IL is visualized to carry out for the benefit of the Indian languages and people. The immediate goals of LDC-IL may be summarized in the following manners:

(a) To become a national repository of linguistic resources for all the Indian languages. In that capacity, it wants to digitize and document all the Indian languages in the form of text, speech, and lexical corpora.
(b) To facilitate the creation of linguistic databases of different forms and types by different Indian organizations that can contribute toward the enrichment of the main LDC-IL repository.
(c) To set up appropriate pan-Indian benchmark standards for data collection, annotation, and storage of electronic corpora for different research and development activities.
(d) To support different language technology development and sharing projects in the Indian language for data collection, processing, and management.
(e) To facilitate training in technical and process-related issues to develop necessary human resources in these areas through workshops, seminars, schools, etc.

(f) To create and maintain LDC-IL Web-based services that will be a gateway for accessing its resources.
(g) To design and provide help in the creation of appropriate language technology tools and systems based on linguistic resources and data for mass utilization, and
(h) To provide a general platform for interactions among the academic institutions, individual researchers, and the masses.

The LDC-IL is quite young in age, but the load it aspires to carry on its shoulder is quite heavy. So there are definitely some doubts if the consortium will be able to perform on the same scale as expected in its missions. This is impossible to predict at this present state of the consortium. But by the way it has begun and the way it functions, we are hopeful that it will grow strength to strength through its dedicated service to the nation and her people over the years.

14.6 Some Other Resources

It is found that during last few years, digital corpora and lexical database of different forms, types, and textures in the Indian languages are also developed and utilized at various levels at different parts of the country. Besides these five major projects related to language corpora and lexical resource generation initiated and funded by the Government of India (elaborated in this chapter), there are also some other projects relating to these areas and these are mostly carried out by public and private sectors at organizational and the state levels, such as the *Microsoft Research India* and the *Society for Natural Language Technology Research*, Government of West Bengal.

Researchers on the Indian languages are not confined to India only, as tools, systems, and resources for some of the Indian languages are also developed in other countries. For instance, research on Bangla is done in Bangladesh and America; on Tamil in Sri Lanka, Singapore, and Malaysia; on Hindi in UK and USA; on Urdu in Pakistan and Germany; and on Punjabi in Pakistan and Canada. Further details of these works are not presented here as these are beyond the defined scope of the present chapter. However, their importance as notable contributions to the Indian language technology development is not ignored here. Since public access to these resources is more or less restricted, it is not possible to retrieve language data and information from these sites and utilize these resources in various works of language processing and technology.

Outside India, there is a center from where we can collect resources of Indian languages—although in a small amount. It is the *Linguistic Data Consortium* of the *University of Pennsylvania* from where we can access the Indian language data and resources, such as *Hindi WordNet* (LDC2008L02), *OGI Multilanguage Corpus* (LDC94S17), *POS Tagset for Hindi* (LDC2010T24), *POS Tagset and tagged corpus of Bengali* (LDC2010T16), *POS Tagset for Sanskrit* (LDC2011T04), and *English Dictionary of Tamil Verb* (LDC2008L01). These resources are developed by various

organizations across the world and are stored in the *Linguistic Data Consortium*, USA, for general access. These are indeed of high relevance in the present context of resource generation and technology development for the Indian languages. As these resources are made publicly available for general utilization and reference, these may be cited as good contributions for the growth and development of the Indian languages and their people.

14.7 Conclusion

With regard to application relevance of these corpora, it may be reported that the TDIL text corpus has been freely downloaded by more than hundred researchers across the world; the ILCI corpora are used by many research organizations in India and abroad; and the *Indian Languages Part-of-Speech Tagging* (ILPOST) tools and corpus of the LDC-IL are also used by many researchers for finalizing the POS tagset for the Indian languages. The ILPOST tagset is essential for cross-verification and comparison of the POS tagset defined within it as well as for the POS tagset defined in the BIS tagset.

There is no doubt in the observation that in the present world of computer-controlled life and living, the generation of digital corpora and tools will not only pave in many new directions of language use in E-governance, E-learning, and online education, but also will present many new findings from language corpora to modify the existing theories, beliefs, and resources of the languages.

Although the works of language corpora generation, lexical resource generation, software and tool development have been a delayed enterprise in India during last few years, at present, it has gathered some momentum through several collective ventures triggered from four major sectors: governmental, institutional, individual, and corporate. If this trend continues for a few more decades, then we can surely expect the emergence of a new India, which is ready with possible solutions to the problems and challenges we face in language technology, information technology, and knowledge engineering.

References

Atkins, S., J. Clear, and N. Ostler. 1992. Corpus Design Criteria. *Literary and Linguistic Computing* 7 (1): 1–16.

Atwell, E., G. Leech, and R. Garside. 1984. Analysis of the LOB Corpus: Progress and Prospects. In *Corpus Linguistics*, ed. J. Aarts and W. Meijs, 41–52. Amsterdam: Rodopi.

Bansal, A., E. Banerjee, and G.N. Jha, 2013. Corpora Creation for Indian Language Technologies—The ILCI Project. In *Proceedings of the 6th LTC (Human Language Technologies as a Challenge for Computer Science and Linguistics)*, 253–257, Poznan, Poland.

Biber, D. 1986. Spoken and Written Textual Dimensions in English. *Language* 62 (4): 384–414.

Dash, N.S. 2007. Indian Scenario in Language Corpus Generation. In *Rainbow of Linguistics*, vol. I, eds. N.S. Dash, P. Dasgupta, and P. Sarkar, 129–162. Kolkata: T. Media Publication.

Dash, N.S., S. Arulmozi, and M. Hussain. 2016. The Carriage of Indian Languages Corpora: And Miles to Go Before We Stop. *Indian Journal of Applied Linguistics* 42 (1 & 2): 63–92.
Francis, N., and H. Kucera. 1964. *Manual of Information to Accompany A Standard Corpus of Present-Day Edited American English*. Providence, R. I.: Deptartmant of Linguistics, Brown University.
Halliday, M.A.K. 1989. *Spoken and Written Language*. Oxford: Oxford University Press.
Hanks, P. 2004. Corpus Pattern Analysis. In *Proceedings of the 11th EURALEX International Congress*, Lorient, France, 6–10 July 2004, eds. G. Williams and S. Vessier, 87–97. Lorient: Universite de Bretagne Sud.
Hanks, P., and J. Pustejovsky. 2005. A Pattern Dictionary for Natural Language Processing. *Revue francaise de linguistique appliqué* 10 (2): 63–82.
Jha, G.N. 2010. The TDIL Program and the Indian Languages Corpora Initiative (ILCI). In *Proceedings of the 7th International Language Resource and Evaluation Conference* (LREC-10), 982–4985, Valletta, Malta, 19–21 May 2010.
Leitner, G. 1991. The Kolhapur Corpus of Indian English: Intravarietal Description and/or Intervarietal Comparison. In *English Computer Corpora: Selected Papers and Research Guide*, ed. S. Johansson and A.-B. Stenström, 215–232. Berlin: Mouton de Gruyter.
Marco, S. 2006. Collocations in Indian English: A Corpus-Based Sample Analysis. *Anglia* 124 (2): 276–316.
Mukherjee, J. 2002. Norms for the Indian English Classroom: A Corpus-Linguistic Perspective. *Indian Journal of Applied Linguistics* 28 (2): 63–82.
Murthy, B.K., and W.R. Deshpande 1998. Language Technology in India: past, present, and the future. In *Proceedings of the SAARC Conference on Extending the Use of Multilingual and Multimedia Information Technology* (EMMIT'98), Pune.
Shastri, S.V. 1988. The Kolhapur Corpus of Indian English and Work Done on Its Basis so Far. *International Computer Archive of Modern English Journal* 2: 15–26.
Shastri, S.V. 1992. Opaque and Transparent Features of Indian English. In *New Directions in English Language Corpora: Methodology, Results, Software Developments*, ed. G. Leitner, 263–275. Berlin: Mouton de Gruyter.
Shastri, S.V. 1996. Using Computer Corpora in the Description of Language with Special Reference to Complementation in Indian English. In *South Asian English: Structure, Use, and Users*, ed. R.J. Baumgardner, 70–81. Urbana, IL: University of Illinois Press.
Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
Wilson, A. 1992. The Usage of 'Since': A Quantitative Comparison of Augustan, Modern British and Modern Indian English. *Lancaster Papers in Linguistics,* no. 80.

Web Links

http://icame.uib.no/kol-eks.html.
http://khnt.hit.uib.no/icame/manuals/kolhapur/index.html.
http:www.tdil-dc.in.
http://www.ldcil.org.
http://research.microsoft.com/en-us/labs/india.
http://www.nltr.org.
http://lrwiki.ldc.upenn.edu.
http://www.ldc.upenn.edu.
http://www.isical.ac.in/~lru/downloadCorpus.html.

Chapter 15
Corpus and Future Indian Needs

Abstract In this chapter, we first try to present a general picture about the present scenario of corpus generation in the Indian context with an appropriate focus on the works already done as well as adequate attention on the works that are in the process of continuation. Along with the reference to text corpora, we also talk about the speech corpora so far developed in a few Indian languages. Moreover, we suggest for generating annotated text and speech corpora in all major Indian languages keeping the applicational relevance of these corpora in various domains of general linguistics, applied linguistics, and language technology. We also argue for generating special corpora in written and spoken texts for exploring their special linguistics features and propose for generation of dialect corpora in all local and regional varieties for their protection and promotion. Finally, we propose for the formation of a national archive or a digital data center for preservation and distribution of Indian text and speech corpora for the benefit of the languages and their speakers.

Keywords Realization · Speech corpora · Annotated corpora · Special corpora Dialect corpora · Monitor corpora · Comparable corpora · National corpus archive · Indian languages

15.1 Introduction

In this chapter, we make a short assessment on the present scenario of Indian activities relating to corpus building in Indian languages as well as focus on the future direction of the whole enterprise. We keep the ardent needs of the future generations of India and other countries into our vision with a realization of the fact that the future course of activities in the fields of linguistics, applied linguistics, language technology, and speech technology will invariably depend heavily on data and information obtained from corpora of actual written and spoken text varieties (Dash 2006). Although the present situation does not support this vision, as no sincere effort is found to be initiated for developing well-formed text and speech corpora in Indian languages,

N. S. Dash and L. Ramamoorthy, *Utility and Application of Language Corpora*,
https://doi.org/10.1007/978-981-13-1801-6_15

we are quite optimistic about it with a focus on some of the ongoing individual and collective efforts made at various parts of the country in this direction (Dash 2008).

The chapter is organized as follows. In Sect. 15.2, we record our impressions and realizations about the state and status of language corpus in the realm of linguistics and language technology; in Sect. 15.3, we inform about the speech corpora so far developed in a few Indian languages; in Sect. 15.4, we argue for generating annotated text and speech corpora in all major Indian languages; in Sect. 15.5, we propose for generating special corpora in both written and spoken texts for exploring their special linguistics features; in Sect. 15.6, we propose for generating dialect corpora in all local and regional varieties for their protection and promotion, in Sect. 15.7, we advocate for development of monitor corpora in all major Indian languages included in the 8th Schedule of the *Constitution of India*; in Sect. 15.8, we suggest for the generation of comparable corpora among the genealogically related Indian languages for initiating cross-linguistic and cross-cultural research; in Sect. 15.9, we propose for the formation of a national archive or data center for preservation and distribution of the Indian text and speech corpora for the benefit of the country; and in Sect. 15.10, we try to show the line of difference between the advanced countries and India with regard to corpus generation and utilization.

15.2 The Realization

From the discussion presented in earlier chapters, it is almost clear to us that language corpora are simply indispensable in language description, building linguistic resources, understanding language communities, as well as in developing systems, tools, and techniques for language and speech technology. The time is now ripe for establishing an intimate functional interface between language and computer so that computer becomes maximally useful for various linguistic tasks for the benefit of the humanity at large. Since a computer cannot execute an independent linguistic work, however small and trivial it may be, it needs adequate linguistic knowledge for its training and execution. If it is properly trained and developed with proper linguistic input it becomes perceptively robust and efficient to do many complex linguistic tasks with ease and accuracy with least human intervention and interference.

However, to train a computer to perform even some routine and not-so-intelligent linguistic tasks is indeed a complex process that asks for proper analysis and valid interpretation of the huge amount of multifaceted and authentic language data compiled in the form of language corpora in digital form. Until and unless an adequate amount of linguistic data, examples, and information are analyzed and provided to a computer for its several trials, training, and simulations, it will miserably fail to execute even a few simple linguistic tasks entrusted to it. This implies that language corpora are simply indispensable not only in the areas of language technology but also in other areas of linguistics as corpora are the most faithful source wherefrom all these domains of analysis and application can heavily draw necessary language data and information. Perhaps, we have no other alternative but to concentrate on the

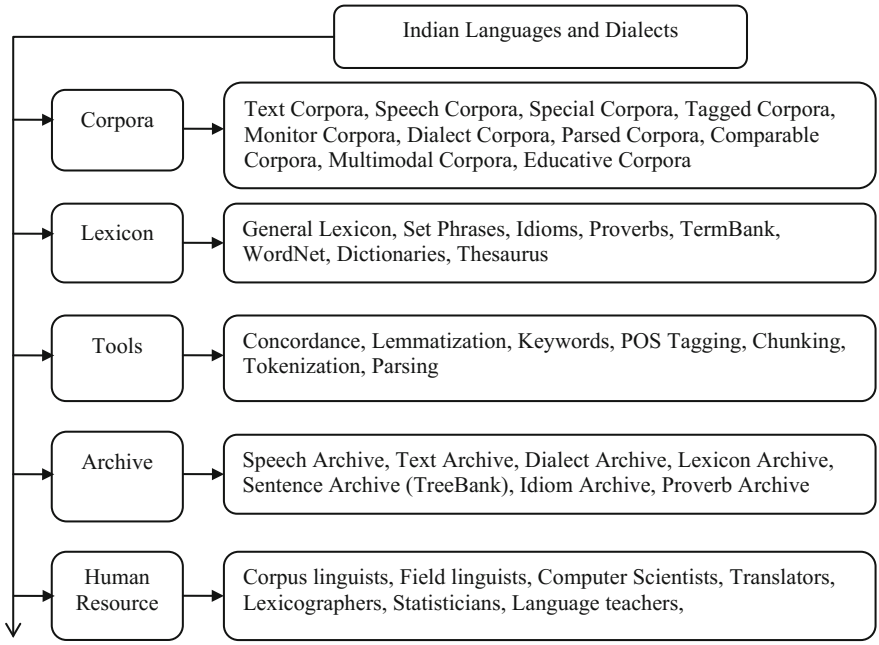

Fig. 15.1 Corpora and others: present and future Indian needs

work of developing corpora of various types in all natural languages to address the needs of linguistics at large.

The present research and development activities in various domains and subdomains of theoretical linguistics, descriptive linguistics, applied linguistics, and language technology in Indian languages invariably need corpora made with samples of written and spoken texts of various size and content, form, and composition. There is also a need for large electronic lexical databases of various types in all major Indian languages developed from corpora for the generation of linguistic resources to be used in mainstream linguistics, language technology, and applied linguistics. At present, we have some written text corpora in some Indian languages developed in the TDIL project, as well as some parallel translation corpora developed in the ILCI project. These are good linguistics resources no doubt, but these are not adequate. Since it is not possible to address all our present requirements with these corpora only, for the sake of future growth and development of Indian languages, we have to take immediate initiatives for developing large, balanced, and adequately representative text and speech corpora in all the major Indian languages. The requirements for the Indian languages are summarized in the diagram presented above (Fig. 15.1).

15.3 Speech Corpora

Nearly three decades ago, a humble attempt was made at the *Indian Statistical Institute* (ISI), Kolkata for collecting some samples of actual speech data for some research works in speech technology (Datta Majumder and Ganguli 1987). After analysis of that speech data, scientists were able to develop software for automatic speech synthesis in Bangla (Ganguli et al. 1988). With the help of this software, the scientists were to develop a tool for speech synthesis in 1990, which was widely known as *the Bangabani*, which could generate synthetic speech in Bangla, Hindi, and in some other Indian speeches (Dutta et al. 1991). The system attained international standard and was successfully deployed to synthesize some Bangla songs, including some songs composed by Rabindranath Tagore, the Nobel Laureate. However, after some years, due to some logistics, financial, technical reasons, this work had come to an end without any chance for reincarnation.

This state has been continued for several years before the whole enterprise was rejuvenated with the active participation of a new group of Indian scholars coming from computer science, linguistics, and information technology. It is heartening to know that speech technology works in Indian languages are gaining momentum in the new millennium with wide expansion of activities relating to various domains of artificial intelligence, cognitive linguistics, and fuzzy logic. Many research institutes and universities of India along with several IT companies are now devoting more time, money, and energy to this area with pre-defined goals for making lasting contributions for the benefit of the languages of the country.

In last two decades, a few Indian institutes have made efforts to develop speech databases from real-life speech events for object-oriented applications. For instance, Patil and Basu (2004) developed a small speech database of Marathi and Hindi speech with samples of dialects spoken at six different regions in Maharastra for the purpose of developing an automatic speaker recognition system. The database carefully includes individual responses made against the questionnaire containing only five questions relating to isolated words, digits, combination-lock phrases, read sentences, and contextual speeches in which each sample contains texts of nearly 90 s. The contextual speech comprised text samples describing nature, memoirs of life, and family life narrated by around 200 local informants. Although this speech database claims high functional and referential relevance within the frame of a specific research goal for which it is developed, it is not a 'speech corpus' in the true sense of the term. It complies following traditional means and methods of speech data collection normally practiced in field linguistics (Samarin 1966). The method used for informant selection, the manners of text sampling, the method of data collection, and the process of text representation in the database are characteristically different from the techniques and methods normally followed in natural speech corpus generation in modern corpus linguistics and language technology.

The importance of speech corpora can be visualized if we try to realize the fact that we need to collect, organize, and disseminate linguistic information of the lesser-known Indian language varieties, many of which are threatened for their eternal

extinction. Therefore, we should initiate projects for speech corpora generation in the Indian languages for proper documentation of the linguistic materials for our social, cultural, and national needs. The lack of joint enterprise in this direction has been one of the bottlenecks for the growth of speech-based research and application in Indian context for several years now (Dash 2003). Moreover, technical difficulties involved in speech corpus generation and research have an impact in a serious manner for blocking its smooth progress in this country. Perhaps, the time has come to direct our attention, fund, and energy in this direction.

For providing a direction, let us enumerate the present Indian needs with regard to speech corpora that may be used in future works of descriptive linguistics, applied linguistics, and speech technology.

(a) We should first develop large, representative, multidimensional, and balanced speech corpora for all major Indian languages. These corpora will contain samples of spoken texts obtained from various real-life speech events. These will be utilized as the basic linguistic resources for various research and development works of general speech description and speech technology for the Indian languages.
(b) We should also generate both synchronic and diachronic speech corpora to identify the newly coined words, terms, and phrases; trace dates of coinage of new terms; identify the patterns of usage of various lexical items in speech; trace the change in meaning of words in spoken usage; observe the changes in the process of sentence construction in verbal mode of expression; and identify the special features of spoken texts that make them different from written texts.
(c) We must have provisions for continuous up-gradation of data in speech corpora so that these become quite useful to reflect on the internal changes which are actually taking place in the speech patterns of people as well as on the life and living of the speakers.
(d) We also require diachronic speech corpora of various types to investigate the course of change observed in the spoken form of a language within a wide range of situations.
(e) We can use both diachronic and synchronic speech corpora to develop dictionaries and grammars of spoken texts as well as to identify variations of speech patterns across spatiotemporal dimensions.

Since the mere collection of speech data and compilation of speech corpora cannot suffice the present needs of speech research and development in the country, there is a strong urge for converting speech corpora into spoken corpora with appropriate annotation. It is acknowledged that spoken corpora have greater functional relevance than speech corpora since spoken corpora usually carry certain additional linguistic information (particularly in the form of annotation) which is not present in raw spoken texts (Dash 2005: 8). The text samples in raw spoken corpora are presented in written form with the addition of some symbols used in phonetic transcription. There are several ways and levels of prosodic annotation, which are often used with transcribed spoken texts to generate spoken corpora (de Ginestel-Maitland et al. 1993; Knowles 1994; Izre'el et al. 2001; Harry 2003). The spoken corpora are useful for

those scholars who do not have adequate technical expertise and skill for analyzing recorded speech data for their works (McEnery and Wilson 1996: 26). Also, these spoken corpora become beneficial for language learners in classroom teachings about the normal speech patterns of the native speakers.

15.4 Annotated Corpora

The present research activities in the area of language technology in the Indian languages require a large amount of annotated text corpora. Therefore, to address this requirement we should immediately develop annotated corpora in the Indian languages. Usually, an annotated version of a corpus contains annotations relating to extralinguistic and intralinguistic information of a language and its users. The extralinguistic annotation normally (but not typically) contains information relating to the following types:

(a) Types of texts,
(b) Year of publication of texts,
(c) Name(s) of sources of texts,
(d) Age information of writer(s),
(e) Gender information of the author(s),
(f) Domain information of use of texts, and
(g) Register information of texts.

The intralinguistic annotation, on the contrary, usually carries information relating to the analytical marks used in a text, parts-of-speech marks of words and phrases, lexical category marks for words and other lexical items, information about the types of sentences, terminal marks for sentences and phrases, semantic information of words, idioms, and phrases, anaphoric information of pronominal forms, identification marks of proper nouns and named entities, and discourse information of texts, etc. The annotated corpora become highly referentially useful for descriptive analysis of a language, studying variations of language use across text types, understanding variations of meanings of words and other elements of a language, observing patterns of lexical change across several decades, and deciphering ambiguity both at the lexical and sentence levels, etc.

In language technology works, annotated text corpora are heavily used for designing tools and systems for morphological processing; decomposing multiword units; generating valid words; parsing sentences; developing electronic lexical databases and TermBanks; compiling electronic dictionaries, grammars, and thesauri; designing systems for machine learning; developing machine-aided translation systems; building automatic text information retrieval systems; for developing automatic language understanding systems, etc., for a natural language.

On the other hand, an annotated speech corpus, although carries tags for both extralinguistic and intralinguistic information of a language, is of a different kind. The process of annotation of a speech corpus varies to a great extent from that of a text

corpus, since a speech corpus is characteristically different from a text corpus (Dash 2006). Information relating to the fields of speech events is used in the description of normal speech as well as in the works of speech technology to explore the nature of context-based variation of speech, semantic, and lexicological changes occurring in spoken interactions and lexicosyntactic ambiguities evoked in speech events.

The extralinguistic annotation in a speech corpus carries information relating to the types of spoken texts, year of text collection, demographic information of speakers, context-based information of text types, source and register variations of text samples, and similar other information. The intralinguistic annotation, on the other hand, generally carries information relating to the analytical marks used in spoken texts, parts-of-speech tagging of words, lexical as well as semantic categories of words, sentence types, discourse information, etc.

Within speech technology research, an annotated speech corpus is considered indispensable for designing tools and systems for speech identification and processing, spoken sentence identification and processing, speech synthesis, speech analysis, speaker identification, speech-to-text conversion, and similar other works. Within the domains of applied and cognitive linguistics, on the other hand, an annotated speech corpus is used for analyzing discourse structures, understanding negotiations, and mediations, interpreting formal and informal conversions, and for understanding the ethnography of verbal communication, etc.

15.5 Special Corpora

There is a strong need for designing special text corpora by way of including samples of written texts composed by different groups of people living in a society (i.e., women, infants, children, students, teenagers, non-native speakers, linguistically impaired people, etc.). For instance, special text corpora may include samples of texts full of jargons and domain-specific terms and codes; samples of texts from private diaries, personal letters, wills, gambler's notes, share market reports, medical reports, auction statements, advertisements, and sports reports to mark the 'special features' of the language varieties used in specific fields of trade, action, and profession. Such text corpora will obviously vary in composition and size according to their patterns of construction and the purpose of their usage. In most cases, special text corpora, due to their unique origin and composition, become highly useful for linguistic investigation and interpretation, since they contain authentic data collected from the texts belonging to particular trades, professions, and activities.

Although most of these special text corpora are not considered fit to contribute directly to the general description of a language, these are, because of their high percentage of unusual linguistic features, considered valuable for projecting into the 'other side' of a language colored with unique linguistic features rarely observed within a normal 'standard language.' However, here is a note of caution: since special text corpora are not properly balanced in composition, so if these are used for the general description of a language, these will definitely project a distorted image of

the standard form of the language. Even then, the basic advantages of special text corpora lie in their text samples, which are collected in such a way that the phenomena we are looking for occur much more frequently in them than found in a general text corpus. Thus, the importance of special text corpora in special linguistic studies is widely attested due to their uniqueness in representing the language variety used by the different classes of people occupying vital segments of a society.

Similar to special text corpora, there is also a need for special speech corpora in the Indian languages. These should include samples of spoken text produced by people belonging to different spheres of life (e.g., infants, learners, teenagers, non-native speakers, women, linguistically impaired people, etc.). Similarly, special speech corpora may be developed with samples of speech data produced by people involved in share market, gambling, auction, medical science, parliamentary debate, court proceeding, underworld activities, etc. The analysis of such speech corpora will show the 'specialities' of a kind of speech used in verbal dialogic interactions in different trades and professions to deal with the co-members working in the same fields of trade, action, and profession.

When compared to general speech corpora, special speech corpora should vary in form, content, and composition according to their patterns of formation and purposes of construction. In some special situations and contexts, they may be considered reliable because they contain spoken texts collected from highly specialized domains. In a reverse way, they contribute to the overall description of speech of a community because text samples will represent a type of speech used by special people, in special situations, and for special purposes. The referential value of special speech corpora in general linguistic description cannot be ignored since they are intentionally designed and developed to represent the special speech varieties used by special classes of people occupying vital positions in the speech community. We should develop both special text corpora and special speech corpora in most of the Indian languages and use them in specific linguistic studies so that we can have better insight about the 'special' use of language in various situations by the people belonging to special categories based on ethnicity, profession, age, gender, linguistic disability, and other sociocultural variables.

15.6 Dialect Corpora

Keeping in view the dialectal diversities observed in the Indian subcontinent, we strongly argue for generating corpora for all the dialects and minority language varieties found in the country. We also argue that systematic collection, digitization, analysis, and documentation of large text samples stored in dialect corpora will help us to identify the total number of speech varieties used in the country, mark similarities and differences between the dialects, investigate power of hegemony playing among the dialects, and preserve their unique linguistic identities both in paper and reality.

In general, dialect corpora should contain elaborate text samples of dialects used in regular dialogic interactions, impromptu conversations, and informal–formal talks

taking place among the people of particular geographical regions. All these features are considered valuable for projecting into the core of the dialects as well as reflecting into the external and internal distinctive features of the dialects and their speakers. Similar to the corpora of standard speeches, all dialect corpora should be properly tagged with various prosodic features following the standards and norms accessed in the annotation of standard speech corpora.

In a multidialectal country like India, where several dialects and minority languages are on the verge of their extinction, we can easily identify multipurpose uses of dialect corpora as essential resources for procuring knowledge from different dialect communities; preserving their language, culture, history, and heritage; developing lexical and syntactic databases from different dialects; developing dictionaries and grammars for dialects; and generating text materials for the dialects for direct use in formal and informal teaching for the benefit of target dialect communities.

15.7 Monitor Corpora

The text corpora developed so far in the Indian languages should be converted into monitor corpora with a system for adding new varieties of text samples. These samples may be obtained from various genres, disciplines, and registers time to time across the regions. Similar to the *British National Corpus*, the *American National Corpus*, the *German Language Corpus*, the *Japanese Text corpus*, etc., these corpora should have provisions for their regular growth and up-gradation so that they become fully fit to reflect on the changes taking place in the Indian languages and their societies. Over the decades, gradually, these corpora will achieve a diachronic dimension to represent the languages fretted with subtle linguistic changes across the generations.

Such diachronic monitor corpora should be designed to cover a wide range of time span to be used as the most trustworthy linguistic resource for identifying newly coined words, terms, and phrases; dates of their coinage; variations in use of linguistic properties; change in meaning of words and lexical items; change in patterns of formation of phrases, idioms, proverbs, and sentences; trace the patterns of change of social psychology as reflected in the languages, etc.

If we are able to develop good monitor corpora for most of the Indian languages, we can be sure that the intuition of native language users, assumptions of linguistic experts, and prophetic speculations of linguistic demigods can be challenged with diachronic referential relevance and verifiable authenticity of monitor corpora in the act of linguistic discovery and validation.

15.8 Comparable Corpora

There should be an effort for developing comparable corpora among those Indian languages, which exhibit close genealogical proximities or typological similarities (e.g., *Hindi–Kashmiri–Punjabi–Urdu*, *Gujarati–Marathi–Konkani*, *Assamese–Bangla–Odia*, *Telugu–Tamil–Kannada–Malayalam*, etc.). Adopting the same process and methods of text samples collection, comparable corpora may be developed with a collection of 'similar' texts in two or more language varieties with an equal amount of data and number of samples taken from each language so that they become maximally comparable among themselves (Landau 2001: 342). In a multilingual country like India, the functional utilities of comparable corpora may easily be visualized in development of bilingual/multilingual lexicon, compiling bilingual/multilingual dictionaries, developing bilingual/multilingual TermBank, producing bilingual/multilingual translational equivalents, writing of bilingual/multilingual grammars, textbooks and reference aids, starting interlingual education and training, and initiating cross-lingual as well as cross-cultural research and development works.

In this context, the referential value of parallel and aligned corpora (e.g., *Malayalam–Tamil*, *Punjabi–Hindi*, *Hindi–Urdu*, *Odia–Bangla*, etc.) of Indian languages is understandable. Similar to comparable corpora, these are valuable resources for devising tools for machine-aided translation systems for the related languages. The application utility of parallel and aligned corpora may easily be envisaged in cross-linguistic research and development, language teaching and training, interlingual communication and information exchange, etc.

In a similar fashion, the development of parallel and aligned corpora between English and the Indian languages (e.g., *English–Hindi*, *English–Bangla*, *English–Telugu*, *English–Tamil*, etc.) is required for domain-specific translation as well as for localization of linguistic and extralinguistic information relating to agriculture, climate and weather, tours and travels, medical science, education and entertainment, science and technology, etc., for the people of India.

15.9 National Corpus Archive

Most of the speech and text corpora developed so far in the Indian languages are incidentally (and unfortunately) within the custody of developers or in the closet of distributing agencies. As a result, these resources are beyond the reach of those people who are not directly involved in corpus generation. Due to this reason, although people from different fields of applied linguistics, linguistics, and language technology are interested in utilizing these corpora in their works, they hardly get any chance to use them. In spite of unlimited scope of application, the corpora built in the 'Technology Development in Indian Languages' (TDIL) project of the DeitY, Government of India in the early 1990s, had hardly been utilized in linguistics and language technology. But we know that these corpora

can be quite fruitfully utilized in language teaching and training, dictionary compilation, disambiguation of word senses, discourse analysis, and language acquisition, etc. Moreover, these can be profitably used as resources and test beds in technology development for optical character recognition, machine-aided translation, electronic lexical resource generation, part-of-speech tagging, morphological processing, information retrieval, text-to-speech conversion, and data mining, etc., for Indian languages. Despite so many varied application possibilities and potentialities, these corpora are not yet made open to one and all interested in these corpora. The time has come to make these resources available to interested scholars of India and abroad for the benefit of the Indian languages and people.

This situation leads us to argue for the formation of a national archive or a data center, in the model of the Oxford Text Archive or the Linguistic Data Consortium, for the Indian languages. This will work for collecting all the corpora developed so far in the Indian languages and scattered all around the country. The central body will work for systematic documentation, preservation, annotation, distribution, and utilization of the corpora for the Indian scholars and others for all kinds of research and development work. The immediate mission of the central body will be the followings:

(a) Collecting corpora already developed by Indian institutes, organizations, and individuals,
(b) Housing these resources in its central digital archive or data center,
(c) Taking necessary initiatives for making these resources readily available to people and agencies involved in corpus-based works in both linguistics and language technology.

The basic activities of the proposed central body will be related to maintaining the databases in digitised archives, producing the databases in complete or select formats in CD-ROMs or in similar other devices, arranging for a network for distribution of corpus databases among the organisations and scholars who want to access these for developing language resources and systems for the benefit of people of the country. The functional modalities of the proposed body may be visualized in the following ways:

(1) It will initiate a concerted effort to assemble all text and speech corpora developed so far in Indian languages and are now under the custody of individual developers.
(2) It will bring these national resources under a single platform for proper digital archiving, documentation, processing, as well as for distribution.
(3) It will act as a centrally approved repository of linguistic data and resources in all the Indian languages in the form of text, speech, and lexical database.
(4) It will facilitate and provide required financial as well as technical support for the creation of new language corpora and databases by different individuals and organizations.
(5) It will set some benchmark standards for corpus collection and storage for the Indian languages for various kinds of research and development works.

(6) It will work for monitoring courses of corpus development activities if the benchmarks defined by it are not properly followed in the creation of future corpora in Indian languages.
(7) It will provide technical support for development and sharing of language processing tools and techniques for corpus collection, storing, processing, and management.
(8) It will work to facilitate development and enrichment of competent manpower and expert human resources through regular training, workshops, seminars, etc., in technical as well as the process-related issues.
(9) It will create and maintain websites and homepages which will be used as primary gateways for accessing Indian language corpora (both written and spoken), corpus processing tools, and other languages resources.
(10) It will work for designing and providing help to the industries and research organizations in the creation of appropriate tools, techniques, systems, and software for language and speech technology for common use.
(11) It will work for providing a link between the individual researchers, academic institutions, and masses so that language corpora, as well as linguistic resources developed from these corpora and tools, become accessible to the people of the country.
(12) It will work for collaboration with similar institutes, universities, and research bodies of other countries to exchange linguistic data, information and knowledge through joint individual or institutional research projects, exchange of scholars, academic sharing, etc., for the benefit of Indian languages and their speakers.

We believe a central archive or a data center proposed here will contribute to a great extent for the creation and enlargement of corpora in the Indian languages. We also expect that such a collective enterprise will ensure high quality of corpora to make them compatible with all kinds of work in general linguistics, applied linguistics, and language technology. Besides, the proposed body should have provisions to support to those who are interested in corpus databases for research and development works in the Indian languages as well as for sharing their resources and tools in the interest of the language communities of the country.

It has also been correctly observed that apart from 23 major Indian languages there are hundreds of minor and tribal languages that deserve an equal amount of attention from the linguists for their analysis and interpretation (Singh 2006: 67). In fact, we urgently need to compile good corpora from these languages in digitized version so that the resources are available to interested scholars for carrying out descriptive analyses of the languages long due in the Indian history of linguistics. Also, generation of corpora in minor and tribal languages will provide scopes for initiating comparative and contrastive studies in the languages across the families and groups with regard to their vocabulary, structure, and function. Therefore, based on the availability of the text materials in printed and electronic forms, attention should be properly diverted toward the creation of text and speech corpora in as

many minor and tribal languages as possible to be in tune with the amount of data already developed in some of the major Indian languages.

Perhaps, it is not unfair if we expect that all research organizations, universities, academic institutions, and corporate houses of India engaged in research and development works in the Indian languages should be interested to participate in corpus-building work. Moreover, the interested institutions, organizations, and individuals may be encouraged to participate in the project works for the accomplishment of the mission still remain unattained for the Indian languages. Let us hope that the idea of generating large-scale corpus databases will take a giant leap in the direction of information technology that India is striving hard to make for decades. It is already understood that large quantum of language databases from different text types are the basic ingredients for research and development of language technology and mainstream linguistics.

The issues relating to collection, processing, annotation, analysis, and utilization of language corpora will eventually compel us to involve a number of people coming from the disciplines like sociology, statistics, computer science, mathematics, ecology, ethnology, anthropology, psychology for proper execution of the tasks ahead. Formalities of all kinds are needed for making corpora maximally error-free, computationally compatible, and operationally optimum to be identified as benchmarks and standards. It, therefore, becomes imperative to conclude that we urgently need to form a central archive for creating and storing language corpora in the Indian languages as well as for sharing these resources among the people and agencies working on the Indian languages to build products for use by the common people.

There is another important issue relevant in this context. It is related to designing a well-formed syllabus for masters in corpus linguistics for the new generation of Indian linguists interested to make their contribution in this new and high potential area of linguistics. We may propose for a model syllabus, which may be used as a rudimentary guideline for introducing corpus linguistics both at undergraduate and postgraduate levels in Indian universities and academic institutions. Moreover, this model syllabus may be used as a part of the syllabus of linguistics proposed and supplied by the *University Grant Commission* (UGC), Government of India. The syllabus is relevant in the context when we think for developing a new generation of linguists who will work in this line for the benefit of the Indian languages and people. A discipline can survive and grow only when it is able to prove its relevance in newer social contexts and when it encourages the new generations to adopt new approaches for the continuation of the discipline. Corpus linguistics is not an exception. It has already established its relevance in the present context of linguistic research and application Andor (2004). Now it needs a large number of followers to continue its functionality in this direction.

15.10 The Contrast

The multipurpose use of corpus in advanced countries far exceeds the amount of its use in the Indian languages. Reasons are many. In the earlier years, the most difficult hurdle was the lack of adequate knowledge about the methods of corpus generation, since it was a new thing on the Indian soil. The work of corpus generation started in India much later after careful consideration and assessment of methods used in many advanced countries. The lack of information and the shades of doubt about the relevance and utility of corpus in linguistics has been another hurdle against its growth and expansion in India. As a result, in comparison with many other countries, India still lags behind not only in corpus generation but also in corpus-based linguistic studies and application. The time has come now for redirecting attention toward this new method of language study to rejuvenate the Indian languages with a new lease of life. The present situation in Indian linguistics definitely needs this life-saving elixir for its survival and growth.

It will not be a sensible idea to work at different domains of language technology until various types of corpora in the Indian languages are generated and processed with due importance and vision. In a country like India, this is an urgent need because language-based application-oriented research and technological development are the basic requirements of the changing time to develop man–machine communication systems for each of the Indian language.

At present, even after 25 years of the first effort for corpus development, the total number of the corpus in the Indian languages is very few, and most of these are also beyond the reach of the majority of people due to some technical and legal factors. The ignorance about the presence of these corpora and their non-availability in electronic forms is equally responsible for blocking the path between the corpora and their users. Therefore, it is really a difficult task to show how corpora in the Indian languages are developed and used for research and application at different parts of the country.

15.11 Conclusion

In this chapter, we have tried to argue for generation of corpora in the Indian languages as well as for their storage, processing, and utilization in research, application, and education in mainstream linguistics, applied linguistics, and language technology. In comparison with other countries, India lags far behind not only in corpus generation but also in corpus-based linguistic studies and application. The time has come for redirecting our attention toward this new method of language study to rejuvenate the discipline with a new lease of life. The present situation in Indian linguistics definitely needs this life-saving elixir for its survival and growth. It will not be sensible to deal with linguistics and language technology until various types of corpora in the Indian languages are generated with due importance and vision.

The human language is a natural, efficient, and highly economical means of thought, expression, and communication. It becomes more effective in those situations where people are in the man-to-man communicative interfaces. Making linguistic communication from one place to another through wire or wireless and interacting with machines, computers, and electronic devices we need to process speech and language data to make these maximally understandable and comprehensible to machines. That means we need some devices that are able to contain language data and are competent to process them with near-human perfection. In a country like India, it is an urgent need because language-based research and technological development is the basic need of the changing time for the purpose of developing man–machine communication systems for each language.

Language technology is highly fruitful to break the language barriers by automatically translating and transmitting information from one language to another. This makes communication easier for people of various linguistic backgrounds. The development of successful and user-friendly devices of language technology, however, requires advanced knowledge from linguistics, acoustics, computer science, information technology, communication technology, signal processing, artificial intelligence, and statistics—all combined in a harmonized fusion for the goals ahead. And this mission visualizes language corpora as the most reliable resource on which the activities of language, linguistics, and technology can depend for definite success to come to the service of science and humanity.

References

Andor, J. 2004. The Master and His Performance: An Interview with Noam Chomsky. *Journal of Intercultural Pragmatics* 1 (1): 93–111.

Dash, N.S. 2003. Corpus Linguistics in India: Present Scenario and Future Direction. *Indian Linguistics* 64 (1–2): 85–113.

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

Dash, N.S. 2006. Speech Corpora vs. Text Corpora: The Need for Separate Development. *Indian Linguistics* 67 (1–4): 65–82.

Dash, N.S. 2008. Corpus Linguistics: An Empirical Approach for Studying a Natural Language. *Language Forum* 34 (2): 5–21.

Datta Majumder, D., and N.R. Ganguli. 1987. Speech Processing Research in India—Perspective and Trends. In *Advances in Computing and Humanities*, vol. I, ed. E. Nissan, 115–159. Connecticut: JAI Press Inc.

de Ginestel-Maitland, A., M. De Calmés, and G. Pérennou. 1993. Multi-level Transcription of Speech Corpora from Orthographic Forms. In *Proceedings of the 3rd European Conference on Speech Communication and Technology (Eurospeech-93)*, vol. II, 1441–1444, Berlin, Germany, 21–23 September 1993.

Dutta, A.K., N.R. Ganguli, and B. Mukherjee. 1991. Nasalisation in Bengali Speech Sounds: Acoustic Phonetic Study. In *Proceedings of 2nd European Conference on Speech Communication and Technology*, vol. 1, 157–180, Geneva, Italy.

Ganguli, N.R., A.K. Dutta, and B. Mukherjee. 1988. Acoustic Phonetics of Non-nasal Standard Bengali Vowels: A Spectrographic Study. *Journal of the IETE* 34 (1): 50–56.

Harry, B. (ed.). 2003. *Corpus Linguistics and Modern Hebrew*. Tel Aviv: Tel Aviv University Press.
Izre'el, S., B. Harry, and G. Rahav. 2001. Designing CoSIH: The Corpus of Spoken Israeli Hebrew. *International Journal of Corpus Linguistics.* 6 (1): 171–197.
Knowles, G. 1994. Annotating Large Speech Corpora: Building on the Experience of MARSEC. *Hermes* 1: 87–98.
Landau, S.I. 2001. *Dictionaries: The Art and Craft of Lexicography*, 2nd ed. Cambridge: Cambridge University Press.
McEnery, T., and A. Wilson. 1996. *Corpus Linguistics*. Edinburgh: Edinburgh University Press.
Patil, H.A., and T.K. Basu. 2004. Multilingual Speech Corpus Design for Speaker Identification in Indian Languages. In *Proceedings of the International Workshop on Standardization of Speech Databases* (Oriental COCOSDA 2004), 8–13, Noida, New Delhi, 17–19 November 2004.
Samarin, W.J. 1966. *Field Linguistics*. New York: Holt, Rinehart, and Winston.
Singh, U.N. 2006. Proposal to Conduct the New Linguistic Survey of India. In *Proceedings of the 28th All India Conference of Linguists* (AICL-28), 22–117, 2–4 November 2006. Varanasi, India: Banaras Hindu University.

Web Links

http://www.linguistics.ucsb.edu/research/santa-barbara-corpus.
http://www.aclweb.org/anthology/Y10-1088.
http://www.cis.upenn.edu/~ccb/publications/arabic-dialect-corpus-2.pdf.
http://corpora.lancs.ac.uk/clmtp/1-data.php.
http://www.ilc.cnr.it/EAGLES/corpustyp/node19.html.
http://www.statmt.org/survey/Topic/ComparableCorpora.
http://www.ilc.cnr.it/EAGLES/corpustyp/node21.html.
http://ota.ox.ac.uk/.
https://www.ldc.upenn.edu/.

Author Index

N. S. Dash and L. Ramamoorthy, *Utility and Application of Language Corpora*,
https://doi.org/10.1007/978-981-13-1801-6

Subject Index

N. S. Dash and L. Ramamoorthy, *Utility and Application of Language Corpora*,
https://doi.org/10.1007/978-981-13-1801-6

M

T