CORPUS LINGUISTICS: AN INTRODUCTION

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

Keywords: Corpus, quantity, representation, simplicity, equality, verifiability, augmentation, genre, documentation, database, language, design, application, generation, text-type, time-span, writers, target-user, written corpus, collection, internet, OCR system, word-entry, sanitation, annotation, character, repetition, transposition, management.

Contents

Glossary
Bibliography
Biographical Sketch

Summary

In this paper we have make an empirical attempt to present a general view about corpus linguistics — a comparatively new field of language research and application. More than half a century ago Corpus Linguistics has started its journey as a field complementary to the mainstream general linguistics, artificial intelligence, computational linguistics, and applied linguistics with direct involvement of computer technology in the area of linguistic research and application. Moreover, within last five decades it has evolved as one of the most promising empirical fields of language study that contribute in a handsome manner for multidimensional growth of mainstream linguistics and language technology in general.

In this paper we have discussed the basic aspects of corpus linguistics. In section two, we have described the definition of corpus as proposed by earlier scholars. In section three, we have discussed some of the salient features of corpus focusing on its basic form and nature. In section four, we have defined corpus typology based on genre of text, nature of data, type of text, purpose of design, and nature of application. In section five, we have addressed some issues related to written corpus generation such as the size of corpus, representation of texts, determination of time span, selection of text documents, selection of text writers, determination of target users, etc. In section six, we have described the process used for generating written corpus that involves factors including methods of text selection, process of text data collection, methods of corpus sanitation, etc. In section seven, we have highlighted the practical utilities of text corpus in various fields of mainstream linguistics and language technology. The article, thus, provides a general idea about the new method of linguistic research and application for the new generations to come.

1. Introduction

The study of language both from empirical and intuitive angles has been one of the oldest trends in the history of human civilisation. In the history, we have emphasised to explore the nature of language to understand how linguistic knowledge has played important roles in cognition and communication. Over the centuries, the field of linguistics has evolved through a long process of cognitive enterprise for establishing conceptual links with other branches of human knowledge. At the dawn of the new millennium, it has now taken a new turn to explore how theories about various aspects of human language are attested in evidence of actual language use manifested in multiple ways of linguistic expression of the common people.

This new direction of language investigation has added an extra dimension to the discipline of traditional linguistics. This has been made possible due to the introduction of computer technology that has helped linguistics to grow and evolve with the supply of tools and techniques to accumulate examples of actual language use form various sectors of linguistic exercises use as well as to analyse these databases in newer

perspectives. The introduction of this new approach has contributed in two basic ways to the field of linguistics in general:

(a) It has enabled the linguists to verify if age-old theories and assumptions about the language and language use are worth pursuing, and

(b) It has provided ample scopes for the direct use of linguistic evidences and information in regular works and activities of linguistics and language technology.

Thus, this new trend of linguistic research and application has worked as an elixir for the revival and survival of the age-old discipline, which was suffering severely from lack of direction, diversion, and application for many years.

We have understood that the invention and advancement of computer technology in the last century has eventually added a new dimension to the field of linguistics. In recent times, as a result of this innovation, a comparatively new field, namely, the Computational Linguistics has evolved as an important area of Artificial Intelligence, which aims at looking at language as an essential instrument of human communication directly linked with human cognition.

Corpus Linguistics, as an important area of computational linguistics, plays an important role. It provides large quantities of empirical language databases accumulated in a systematic manner from various fields of actual language use following some statistical methods and techniques of data sampling. Also, it provides some sophisticated devices to analyse these corpora to extract linguistic data, examples, and information necessary in applied linguistics, computational linguistics, and artificial intelligence for understanding human language in a better way as well as for applying this data and information in various fields of human knowledge.

There is always a strong cognitive and linguistic motivation to envisage how we communicate through language across time and space. There is also a technical motivation to build up an intelligent computer system that will be able to make efficient linguistic interaction with human. With these motivations, both computer scientists and linguists, in recent times, have joined hands together to develop systems such as machine translation, information extraction, language understanding and generation, speech understanding and generation, computer-aided language teaching, etc. that contribute for the benefit and advancement of the whole mankind. However, for designing and developing such a system, we need to understand empirically the natural languages adorned with all their regular and rare features. Here, language corpora become indispensable, since they have good potentials to exhibit most of the features of a natural language manifested within large collection of empirical databases.

At present, scientists across the world have been engaged in computerising information of various types, since the primary goals of computational linguistics are to characterise, as far as possible, the features of a natural language within the frame of computer architecture. It has become necessary to carry out investigation into language corpora to benefit from information and insights expressed in linguistic analysis of natural language databases stored within corpora. For instance, if we want to make proper

interpretation of a simple sentence of a language by computer, we need prior information of linguistic analysis of such sentences carried out by experts to empower the system. Thus, description and analysis of linguistic properties stored within corpora become significant inputs in both computational linguistics and applied linguistics. In fact, information obtained from corpora does not contribute to the field of computational linguistics alone. It equally provides valuable insights about the description and understanding of a language — an important part of language description and learning.

2. What is a Corpus?

The term *corpus* is derived from the Latin word *corpus* that means "body". The Latin term, however, displays two distinct descendants in modern English:

(a) *corpse* (it came via Old French *cors*) and
(b) *corps* (it came via modern French *corps* in the 18th century).

The first form (i.e. *corpse*) entered into English in the thirteenth century as *cors* and during the fourteenth century it had its original Latin '*p*' reinserted. At first it meant simply 'body', but by the end of the fourteenth century, the sense 'dead body' became firmly established. However, on the other hand, the original Latin term *corpus* itself was acquired in English in the fourteenth century (Ayto 1990: 138).

Within the domain of modern corpus linguistics, the term 'corpus' refers to *"a large collection of linguistic data, either written texts or a transcription of recorded speech, which can be used as a starting point of linguistic description or as a means of verifying hypotheses about a language"* (Crystal 1995). Thus, it refers to a large collection of written and spoken text samples, available in machine-readable form, accumulated in scientific manner to represent a particular variety or use of a language.

According to scholars, a corpus is a collection of linguistic items that are selected and ordered according to some explicit linguistic criteria defined by the users in order to be used as a sample of a language. It is methodically designed to contain millions of word compiled from diverse text types across many demographic variations to encompass the diversity a natural language exhibits through its multifaceted use. McEnery and Wilson (1996: 215) have classified corpus in a finer scheme of classification characterised by its inherent features:

(a) Loosely, a corpus refers to any body of text;
(b) Most commonly, it refers to a body of machine-readable text; and
(c) More strictly, it refers to a finite collection of machine-readable texts sampled to be maximally representative of a language or a variety of it.

In principle, a corpus is actually designed for accurate study of the linguistic properties, features, and phenomena observed in a language. Therefore, we have argued that a systematically compiled corpus, however small in size, should adhere to the following criteria (Dash 2005: 12):

- A corpus should faithfully represent both the common and special linguistic features

of the language from which it is designed and developed. The idea of text representation in a corpus indirectly refers to the total sum of its components (i.e. words, phrases, clauses, sentences, etc.) included in it. However, in practice, the total number of words included in a corpus may determine its size but may fail to abide by the principle of proper text representation. Therefore it is better to keep fields open for a corpus as well as keep number of words unlimited for the benefit of language and users.

- A corpus should be large and wide to encompass texts from various disciplines. In other words, directional varieties of language use manifested in various disciplines and domains should have proportional representation in it. For instance, text samples from the fields of natural sciences should carry equal weight as those from aesthetics, literature, mass media, engineering, and social sciences. Thus, a balanced representation of text samples obtained from all disciplines and domains of language use will ensure its reliability.

- A corpus should be a true replica of physical texts available in printed form. Thus, it should faithfully preserve various word forms, spelling variations, punctuation marks as well as various other orthographic symbols used in the source texts. Else, the actual image of a language or the language variety will be distorted and a corpus will lose its value and authenticity.

- A corpus should be available in the electronic form for easy access by the end users in order to enable common users as well as language researchers to use the database in multiple tasks related to language description and analysis, statistical analysis, language processing, translation, etc.

As corpus designers our basic task is to gather large amount of representative text samples covering wide varieties of language used in various domains of our regular linguistic interaction. Since a corpus is capable of representing potentially unlimited selections of text, it may be defined acrostically from the letters used to compose the term in following way (Dash 2005: 4):

C : Compatible to both man and computer,
O : Operational in research and application,
R : Representative of a language or a variety,
P : Processable by both man and machine,
U : Unlimited in the amount of data and samples, and
S : Systematic both in formation and representation.

Unless defined otherwise, let us assume that a corpus will possess all the properties mentioned above. Exception may be made for historical corpora, which have limited use due to their diachronic form and composition. Historical corpora are mostly used within specific areas of historical linguistics that attests indirect importance in the field of empirical language research. In essence, a well-defined and systematically developed corpus is an empirical standard, which acts as a valuable benchmark for validation of usage of all linguistic properties available in a natural language.

3. Salient Features of Corpus

A corpus, in principle, is assumed to have certain characteristic features discussed in the following subsections. It implies that a corpus, which possesses one or more non-default values for the characteristic features of a general corpus, may be identified as a 'special corpus' the title of which will specify its deviation from the general frame of a general corpus.

3.1. Quantity

The most common question, which is often raised by the new comers in corpus linguistics, is: how large a corpus do we need to generate? It is not easy to provide an answer to this question by recommending any set of figures. However, the term 'quantity' means that a general corpus shall contain a very large amount of language data either in spoken or in written form. In fact, the size of a corpus is virtually the sum of the size of its components used to constitute its body. The whole point of assembling a corpus is to collect language databases in large quantities although the advent of 'monitor corpus' changes the idea of size from a 'total amount' to a 'rate of flow'.

The question of quantity shall be envisaged within the basic framework of technology development. In the early years of electronic corpus generation, the *Brown Corpus* that contains just one million words was considered to be a standard one. In the *Brown Corpus* one million words were divided evenly into several genres with five hundred samples of text that contained two thousand words each. These samples were obtained from various written and published texts of modern English. By mid-seventies, however, the target of the number of words shoots up by an order of magnitude. As a consequence, the *Birmingham Collection of English Text* ends up with twenty million words in 1985. In the mid-nineties, the *Bank of English* closes with two hundred million words. At the dawn of new millennium, it obtains an exponential dimension with unbridled collection of words from all possible sources of text. For instance, the *British National Corpus* accumulates more than four hundred million words within a span of few years.

The idea of size of a corpus indirectly reflects on the ease or difficulty of acquiring text samples. This is also loosely related to availability of text materials in a language. Generally, the socio-politically influential languages such as English, German, French, Hindi, etc. have easily available text materials. This is, however, not the situation for languages used in less advanced countries of Asia and Africa.

However, regarding the corpus of spoken text samples, the most influential languages as well as least influential languages present almost the same scenario. In both the language types the amount of spoken texts in formal and informal conversation are not easily available even though advanced language enjoy certain facilities hardly accessible to less advanced languages.

3.2. Quality

The default value of quality of a corpus directly refers to its 'authenticity'. That means

language databases shall be gathered from genuine normal spoken and written texts. The basic role of a data collector is confined here within the act of acquiring data from normal texts for corpus generation. (S)he should bear in mind the need of the corpus users to protect the interests of the people who will be using the corpus databases to formulate linguistic statements about the way a language is actually used in normal situations.

A corpus collector, therefore, has no right to include text samples obtained from experimental conditions or/and artificial circumstances. It is indeed difficult to draw a line between the two types of text, because, for instance, texts from television interviews may appear to be natural but these are deliberately put under artificial conditions to get extremely odd responses. On the other hand, normal casual conversations are expected to be impromptu and spontaneous but may be rehearsed by one or more participants for presentation in a discourse.

3.3. Representation

A general corpus, in principle, shall include samples from a wide range of texts in order to attain proper representation of a language. Moreover, the corpus should be balanced to include text samples from all disciplines and subject fields to represent maximum number of linguistic features found in a language. The databases stored in a corpus should be authentic in their representation of the source text, since future linguistic analysis and investigation systems based on the databases will need verification and authentication of information derived from a corpus representing the language in question.

3.4. Simplicity

This particular feature denotes that a corpus should contain text materials in simple and plain text format so that corpus users have easy access to the plain texts without stumbling upon any additional linguistic information tagged up within text samples. At present, there are a few corpora in which text samples are tagged in SGML (i.e. Standard Generalized Mark-up Language, ISO 8879: 1986) format where all mark-ups are carefully used not to impose any additional burden of information on the text samples.

Normally, the role of a mark-up system, in relation to text representation, is to preserve, in linear encoding, some of the linguistic and non-linguistic features, which will otherwise be lost at the time corpus processing. The system of text encoding or annotation is perceived to be highly useful, since its presence enhances easy retrieval linguistic information from corpus.

3.5. Equality

Text samples collected in a corpus should be of equal size with regard to number of words preserved in each sample. However, this is considered to controversial, since it cannot be adopted everywhere in a uniform manner. Size of text samples will vary proportionally depending on the needs of the users as well as on availability of text

materials. At the early stage of electronic corpus generation, the size of the *Brown Corpus* acquired great importance to act as a guideline in the context of generating corpora in other languages. At present, however, the concept is changed considerably to make a corpus both multidimensional and representative.

3.6. Retrievability

Language data stored in a corpus should be easily retrievable and usable by the end-users. Therefore attention should be paid to the techniques used to preserve language data in electronic form in computer or in digital archive for the users. The present technology makes it possible for users to generate corpus in a personal computer and preserve it in such way that anybody can easily retrieve data and information as and when required.

3.7. Verifiability

Text samples collected from various sources must be authentic and reliable in representation of a language under investigation. Without the scope for empirical verification, the importance of a corpus is reduced to zero. In fact, this quality makes a corpus trustworthy, because a corpus becomes accessible for all types of empirical investigation by users either to verify earlier views or to refute existing observations. In fact, this particular feature has put corpus linguistics miles ahead of generative models of language study and investigation.

-
-
-

Bibliography

Atkins, Sue, Jereme Clear, and Nicholas Ostler (1992) “Corpus design criteria”. *Literary and Linguistic Computing*. **7**(1): 1-16. [This is perhaps the first article that gives us solid direction about the criteria to be considered for designing a corpus in a language]

Ayto, J. (1990) *Dictionary of Word Origin.* London: Blumsberry. [A highly useful dictionary about the origin of nearly 2000 English words]

Biber, D. (1993) “Representativeness in corpus design”. *Literary and Linguistic Computing.* **8**(4): 243-257. [it contains valuable discussion on the methods of presenting text samples in a corpus]

Botley, S., A. McEnery, and A. Wilson (Eds.) (2000) *Multilingual Corpora in Teaching and Research*. Amsterdam-Atlanta, GA.: Rodopi. Pp. 177-191. [Perhaps, the first book to deal with the utilisation of language corpora in language teaching – both primary and secondary]

Crystal, D. (1995) *The Cambridge Encyclopaedia of the English Language.* Cambridge: Cambridge University Press. [A highly useful book for referring and understanding various terminologies used in linguistics]

Dash, N.S. (2005) *Corpus Linguistics and Language Technology*. New Delhi: Mittal Publications. [The first book to deal with topics of corpus linguistics and language technology in any Indian language. Also, the most resourceful book that covers almost all aspects of corpus linguistics in general]

Dash, N.S. (2006) "Speech corpora Vs. Text Corpora: Need for Separate Development". *Indian Linguistics*. (D.P. Pattanayak Felicitation Numbers). Vol. 67. Nos. 1-4. Pp. 65-82. [It strongly argues for adopting separate strategies for generation and utilization of text and speech corpora]

Hunston, S. (2002) *Corpora in Applied Linguistics.* Cambridge: Cambridge University Press. [A very useful book that deals with the use of language corpora in language teaching]

Kennedy, G. (1998) *An Introduction to Corpus Linguistics.* London: Addison-Wesley Longman. [One of the introductory books to deal with corpus linguistics. It pays utmost importance to English language corpora developed across the contries]

Leech, G. (1991) "The state of the art in corpus linguistics". In, Aijmer, K. and B. Altenberg (Eds.) *English Corpus Linguistics: Studies in Honour of Jan Svartvik.* London: Longman. Pp. 8-29. [An insightful analysis about the present status of corpus linguistics and its future prospects]

Leech, G. (1992) "Corpus linguistics and theories of linguistic performance". In, J. Svartvik (Ed.) *New Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82*. Berlin: Mouton de Gruyter. Pp. 125-148. [Highly useful for understanding how information and evidence acquired from corpus analysis can change traditional concepts and theories of linguistic performance of users]

Ljung, M. (Ed.) (1997) *Corpus-Based Studies in English: Papers from the Seventeenth International Conference on English-Language Research Based on Computerized Corpora*. Amsterdam-Atlanta, GA.: Rodopi. [A useful book that shows how study of stylistics in English is greatly revised and redesigned with support of data and information obtained from language corpora]

McEnery, T. and A. Wilson (1996) *Corpus Linguistics*. Edinburgh: Edinburgh University Press. [One of the highly useful introductory books on corpus linguistics. Although not all pervasive, yet quite useful for the beginners]

Sinclair, J. (1991) *Corpus, Concordance, Collocation*. Oxford: Oxford University Press. [A milestone in the history of growth and development of corpus linguistics. A must read for all the people working in this area. Quite insightful and directional].

Summers, D. (1991) *Longman/Lancaster English Language Corpus: Criteria and Design.* Harlow: Longman. [Good introductory book on the history of the Lancaster corpus designing. It deals with the criteria and the designing principles which can be used as models for other corpora]

Svartvik, J. (1986) "For Nelson Francis". *ICAME News*. No. **10**: 8-9. [Fantastic speculation on multipurpose use of language corpora]

Svartvik, J. (Ed.) (1992) *New Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82, Stockholm*. Berlin: Mouton de Gruyter. [One of the best books on corpus linguistics ever edited. It is an anthology that includes several articles of corpus linguistics written by experts of the field]

Wichmann, A., S. Fligelstone, A. McEnery, and G. Knowles (Eds.) (1997) *Teaching and Language Corpora*. London & New York: Addison Wesley Longman. [Another useful book that deals with the use of language data, information, and evidence from corpora for language teaching]

Wills, J. D. (1990) *The Lexical Syllabus.* London: Collins. [It deals with an entirely new method of language teaching. It argues that leaning the usage patterns of lexical items will give learners better competence on grammar of a language].

Winograd, T. (1983) *Language as a Cognitive Process.* Vol. I. Mass.: Addison-Wesley. [A gGood book in the area of cognitive linguistics. It deals with problems, possibilities, and models of language processing and understanding by both man and machine]

Biographical Sketch

Niladri Sekhar Dash has been working in the area of corpus linguistics and language technology for more than fifteen years at the *Indian Statistical Institute*, Kolkata. His first book "Corpus Linguistics and Language Technology: With Reference to Indian Languages" (2005: Mittal Publications, New Delhi) is widely acclaimed as one of the most exhaustive works in this area and used as a course and reference

book in several universities and research institutes in India and abroad. He has two more books in this area published in Bengali which are appreciated as first works in Bengali. To his credit Dash has more than fifty research papers published in national and international journals. He has taught as a visiting faculty at *Madras University*, Chennai (India), *Jadavpur University*, Kolkata (India), *Punjabi University*, Patiala (India), and *North Bengal University*, Darjeeling (India). He has acted as a Co-Investigator in the *TDIL Project* of the *Ministry of Information Technology*, Govt. of India, besides acting as an Expert in the ASI@IT&C project of *European Commission*, and in the Indo-African project of the *International Scientific Research Network*, Brazil. Dash specialises in the area of corpus linguistics, language technology, and lexicography.