SCIENCE BOX

Journal of Advanced Research and Stability
Volume: 02 Issue: 12 | Dec - 2022 *ISSN: 2181-2608*
www.sciencebox.uz

CORPUS LINGUISTICS: A GENERAL INTRODUCTION

Ma'ripov Jalolkhan Kamoliddin ugli
Jizzakh branch of the National University of Uzbekistan named after Mirzo Ulugbek, The faculty of psychology, the teacher at the department of foreign languages

Annotation*: Corpus Linguistics is a multidimensional area. It is an area with a wide spectrum for encompassing all diversities of language use in all domains of linguistic interaction, communication, and comprehension. The introduction of corpus in language study and application has incorporated a new dimension to linguistics.*

Keywords: *corpus, linguistics, quantity, quality, representation, simplicity, equality, retrieavability, verifiability, augmentation.*

In principle, Corpus Linguistics is an approach that aims at investigating language and all its properties by analyzing large collections of text samples. This approach has been used in a number of research areas for ages: from descriptive study of a language, to language education, to lexicography, etc. It broadly refers to exhaustive analysis of any substantial amount of authentic, spoken and/or written text samples. In general, it covers large amount of machine-readable data of actual language use that includes the collections of literary and non-literary text samples to reflect on both the synchronic and diachronic aspects of a language.

The uniqueness corpus linguistics lies in its way of using modern computer technology in collection of language data, methods used in processing language databases, techniques used in language data and information retrieval, and strategies used in application of these in all kinds' language-related research and development activities. Electronic (digital) language corpus is a new thing. It has a history of nearly half a century. Therefore, we are yet to come to a common consensus as to what counts as corpus, and how it should be designed, developed, classified, processed and utilized.

The basic philosophy behind corpus linguistics has two wings: (a) we have a cognitive drive to know how people use language in their daily communication activities, and (b) if it is possible to build up intelligent systems that can efficiently interact with human beings. With this motivation both computer scientists and linguists have come together to develop language corpus that can be used for designing intelligent systems (e.g., machine translation system, language processing system, speech understanding system, text analysis and understanding system, computer aided instruction system, etc.) for the benefit of the language community at large.

All branches of linguistics and language technology can benefit from insights obtained from analysis of corpora. Thus, description and analysis of linguistic properties collected from a corpus becomes of paramount importance in all many areas of human knowledge and application.

The term corpus is derived from Latin corpus "body". At present, it means representative collection of texts of a given language, dialect or other subset of a language to be used for linguistic analysis. In finer definition, it refers to (a) (loosely) anybody of text; (b) (most commonly) a body of machine-readable text; and (c) (more strictly) a finite collection of machine-readable texts sampled to be representative of a language or variety (McEnery and

Wilson 1996: 218).

Corpus contains a large collection of representative samples of texts covering different varieties of language used in various domains of linguistic interactions. Theoretically, corpus is (C)apable (O)f (R)epresenting (P)otentially (U)nlimited (S)elections of texts. It is compatible to computer, operational in research and application, representative of the source language, processable by man and machine, unlimited in data, and systematic in formation and representation (Dash 2005: 35).

According to Niladri Sekhar Dash basic features of corpus can be divided into the following types:

- Quantity: It should be big in size containing large amount of data either in spoken or written form. Size is virtually the sum of its components, which constitute its body.
- Quality (= authenticity). All texts should be obtained from actual examples of speech and writing. The role of a linguist is very important here. He has to verify if language data is collected from ordinary communication, and not from experimental conditions or artificial circumstances.
- Representation: It should include samples from a wide range of texts. It should be balanced to all areas of language use to represent maximum linguistic diversities, as future analysis devised on it needs verification and authentication of information from the corpus representing a language.
- Simplicity: It should contain plain texts in simple format. This means that we expect an unbroken string of characters (or words) without any additional linguistic information marked-up within texts. A simple plain text is opposed to any kind of annotation with various types of linguistic and non-linguistic information.
- Equality: Samples used in corpus should be of even size. However, this is a controversial issue and will not be adopted everywhere. Sampling model may change considerably to make a corpus more representative and multi-dimensional.
- Retrieavability: Data, information, examples, and references should be easily retrievable from corpus by the end-users. This pays attention to preserving techniques of language data in electronic format in computer. The present technology makes it possible to generate corpus in PC and preserve it in such way that we can easily retrieve data as and when required.
- Verifiability: Corpus should be open to any kind of empirical verification. We can use data form corpus for any kind of verification. This puts corpus linguistics steps ahead of intuitive approach to language study.
- Augmentation: It should be increased regularly. This will put corpus 'at par' to register linguistic changes occurring in a language in course of time. Over time, by addition of new linguistic data, a corpus achieves historical dimension for diachronic studies, and for displaying linguistic cues to arrest changes in life and society.
- Documentation: Full information of components should be kept separate from the text itself. It is always better to keep documentation information separate from the text, and include only a minimal header containing reference to documentation. In case of corpus management, this allows effective separation of plain texts from annotation with only a small amount of programming effort.

References:

1. Alshawi H. Memory and context for language interpretation. - Cambridge, 1987. - 188 p
2. Anderson J. R. Language, memory, and thought. - New York, 1976. - 291 p.
3. Bell R.T. Translation and Translating Theory and Practice. - London, New York. Longman, 1991. -298 p.
4. Brewka G. Principles of Knowledge Representation. - California, 1996. - 318 p.
5. Charniak E. On the Use of Framed Knowledge in Language Comprehension// Artificial Intelligence, vol.11, 1978. - P.225 - 265.
6. Charniak E. Organization and Inference in a Frame-Like System of Common Sense Knowledge. - Castagnola, ISCS, 1975.
7. Fillmore Ch.J. The Case for Case Reopened // P.Cole, J.M.Sadock. Syntax and Semantics, 8: Grammatical Relations. -N.Y.: Academic Press, 1977. - P.59 - 82.
8. Johnson-Laird Ph.N. Mental Models. - Cambridge (Mass.]. Harvard Univ.Press, 1983.-XIII, 513 p.
9. Klix F. On stationary and infernal knowledge // XXIInd International congress of psychology (Leipzig, GDR, July 6 - 12, 1980) . Abstract guide. - Leipzig, 190. - P. 27 - 29.
10. Lakoff G., Johnson M. Metaphors We Live by. - Chicago; London: Univ. of Chicago Press, 1980.-XIII, 242 p.
11. Rumelhart D.E. Notes for a Schema for Stories// representation and Understanding: Studies in Cognitive Science/ Ed. by D.Bobrow, A.Collins. - N.Y.: Academic Press, 1975. -P.211 - 236.
12. Schank R.C. Dynamic Memory: A theory of Reminding and Learning in Computers and People. - Cambridge etc.: Cambridge Univ. Press, 1982. - XV, 234 p.
13. Schank R.C., Abelson R.P. Scripts, Plans, Goals and Understanding. - Hillside, NJ: Lawrence Erlbaum, 1977. - 248 p.
14. FEATURES OF ANTHROPOCENTRIC STUDY OF SACRED TEXTS. Maripov Jalolkhan Kamoliddin ugli, Alimkulova Khulkar Tolibovna. JournalNX- A Multidisciplinary Peer Reviewed Journal, ISSN No: 2581 – 4230 VOLUME 8, ISSUE 1, Jan. -2022.
15. DINIY TA'LIMOTLAR VA DINIY MATNLARNI TAHLIL

    QILISHNING O`ZIGA XOS JIHATLARI. Ma'ripov Jalolkhan Kamoliddin ugli, Alimkulova Khulkar Tolibovna. «ОБРАЗОВАНИЕ И НАУКА В XXI ВЕКЕ». Выпуск №22 (том 2) (январь, 2022)
16. N. S. Dash: Corpus Linguistics: A General Introduction. CIIL, Mysore, 25th August 2010