IJARIIT INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact Factor: 6.078
(Volume 7, Issue 6 - V7I6-1206)
Available online at: https://www.ijariit.com

Using corpus Linguistics in the development of writing: A Review

Dr. Layth Hussein Mohamdsalih
taifali607@gmail.com
Gujarat University, Ahmedabad, Gujarat

ABSTRACT

Using Corpus Linguistics in the Teaching of Writing Skills Corpus linguistics has a strong justification in the teaching of writing skills because it can identify patterns of authentic language use by analyzing actual usage. The purpose of this study is to list the most frequent corpora and to investigate helpful, economical, and user-friendly software programs used in the development of writing abilities, such as the Word Smith Tools or Text STAT. In addition, the author gives various examples of how corpus linguistics might be used to help students improve their writing skills. For example, the concordance allows you to see any word or phrase in context so you can identify who it hangs around with. As a result, pupils can see the differences between words they frequently mix up (e.g., excited vs. exciting). the emergence of writing.

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*

1. INTRODUCTION

One of the earliest trends in the history of human civilization has been the empirical and intuitive study of language. In the past, we have emphasized the need of studying the nature of language in order to comprehend how linguistic knowledge has influenced cognition and communication. The study of linguistics has evolved over centuries through a long process of cognitive effort aimed at building conceptual connections with other branches of human knowledge. It has now taken a new turn at the turn of the millennium to investigate how ideas about various characteristics of human language are attested in evidence of actual language use exhibited in a variety of linguistic expressions by ordinary people.

Traditional linguistics has gained a new dimension as a result of this new trend in language research. This has been made feasible by the arrival of computer technology, which has enabled linguistics to expand and improve by providing tools and procedures for collecting samples of actual language use from various areas of linguistic activities and analyzing these databases in innovative ways. The introduction of this new approach has benefited the field of linguistics in two ways: (a) it has allowed linguists to test whether age-old theories and assumptions about language and language use are still valid, and (b) it has provided ample opportunities for the direct use of linguistic evidence and information in regular linguistics and language technology works and activities. As a result, this new language study and application trend has served as an elixir for the restoration and survival of an age-old discipline that had been ailing for many years from a lack of direction, distraction, and application. We have recognized that the development and expansion of computer technology over the last century has given linguistics a new dimension. As a result of this advancement, a relatively new discipline known as Computational Linguistics has emerged as an important area of Artificial Intelligence in recent years. Its goal is to examine language as a fundamental instrument of human communication that is directly linked to human cognition. Corpus linguistics, as a branch of computational linguistics, has a significant role to play. It delivers vast amounts of empirical language databases gathered in a systematic manner from diverse domains of actual language use using statistical approaches and data collecting techniques. It also includes sophisticated tools for analyzing these corpora and extracting the linguistic data, examples, and information needed in applied linguistics, computational linguistics, and artificial intelligence to better understand human language and apply this data and information to various fields of human knowledge. There is always a strong cognitive and linguistic motive to consider how humans interact across time and location using language. A technical reason exists to develop an intelligent computer system capable of efficient verbal interaction with humans. With these motivations in mind, computer scientists and linguists have collaborated in recent years to develop systems such as machine translation, information extraction, language understanding and generation, speech understanding and generation, computer-assisted language teaching, and others that benefit and advance humanity as a whole. However, in order to create and construct such a system, we must first have a thorough understanding of natural languages, including all of its common and uncommon properties. Language corpora become indispensable in this situation because they have the potential to demonstrate most of the characteristics of natural language in a wide collection of empirical information.

2. WHAT IS A CORPUS?

The term corpora come from the Latin word corpus, which literally means "body." The Latin term, on the other hand, has two unique English descendants: (a) corpse (from Old French cors) and (b) corpse (it came via modern French corps in the 18th century).

The initial form (i.e., corpse) entered English as cors in the thirteenth century, and its original Latin 'p' was reinserted in the 14 century. It originally meant simply 'body,' but by the end of the fourteenth century, it had taken on the meaning of 'dead body.' The original Latin term corpus, on the other hand, was gained in English in the fourteenth century (Ayto 1990: 138).

The term 'corpus' refers to "a huge collection of linguistic data, either written texts or a transcription of recorded speech, that can be utilized as a starting point for linguistic description or as a means of confirming hypotheses about a language" in the context of current corpus linguistics (Crystal 1995). As a result, it refers to a vast collection of machine-readable written and spoken text samples gathered in a scientific manner to reflect a particular variant or use of a language. A corpus, according to scholars, is a collection of linguistic objects that are picked and ordered based on specific linguistic criteria established by users in order to be used as a sample of a language. It is meticulously built to contain millions of words collated from a variety of text kinds across a wide range of demographics in order to capture the diversity that a natural language demonstrates through its many applications. McEnery and Wilson (1996: 215) categorized corpus using a finer classification approach based on its inherent characteristics:

A corpus is a finite collection of machine-readable texts sampled to be maximally representative of a language or a variety of languages. In theory, a corpus is intended to be used to investigate the linguistic traits, features, and phenomena that are noticed in a language. As a result, we've argued that a methodically compiled corpus, no matter how tiny, should meet the following requirements (Dash 2005: 12):

3. SALIENT FEATURES OF CORPUS

In theory, a corpus should contain specific characteristics, which are addressed in the following subsections. It indicates that a corpus with one or more non-default values for the distinguishing aspects of a general corpus can be classified as a 'special corpus,' the title of which will specify the corpus's departure from the general frame.

3.1. Quantity

The most common question that newcomers to corpus linguistics have is: how large of a corpus do we need to generate? It's difficult to respond to this question by recommending a set of figures. The term 'quantity,' on the other hand, implies that a broad corpus should contain a vast amount of language data, either spoken or written. In fact, the size of a corpus is essentially equal to the total of the sizes of its constituents that make up its body. The goal of building a corpus is to collect huge amounts of linguistic databases, while the introduction of the 'monitor corpus' shifts the concept of size from a 'total amount' to a 'pace of flow.

Within the basic framework of technological growth, the issue of quantity must be considered. The Brown Corpus, which comprises only one million words, was considered a standard corpus in the early days of electronic corpus production. One million words were divided evenly between numerous genres in the Brown Corpus, with 500 samples of text containing two thousand words each. These samples were taken from a variety of written and published modern English writings. However, by the mid-1970s, the aim for the quantity of words had increased by an order of magnitude. As a result, by 1985, the Birmingham Collection of English Text had grown to twenty million words. The Bank of English closes with 200 million words in the mid-nineties. It takes on an exponential dimension towards the turn of the millennium, with an unfettered gathering of words from all imaginable text sources. The British National Corpus, for example, has accumulated over 400 million words in just a few years. The size of a corpus has an indirect impact on how easy or difficult it is to obtain text samples. This has a tangential relationship with the availability of text materials in a language. In general, text resources for socio-political prominent languages such as English, German, French, Hindi, and others are readily available. This is not the case, however, for languages spoken in Asia and Africa's less developed countries.

3.2. Quality

A corpus's default quality value directly refers to its 'authenticity.' That is to say, language databases will be constructed from real-life spoken and written texts. The primary function of a data collector is to collect data from ordinary texts in order to create a corpus. She should keep in mind that corpus users need to protect the interests of others who will be using corpus databases to develop linguistic assertions about how a language is used in everyday contexts. As a result, a corpus collector has no authority to incorporate text samples obtained under experimental or contrived conditions. It is difficult to make a distinction between the two sorts of text since, for example, television interview texts may appear natural, but they are purposely placed under artificial conditions in order to elicit really strange reactions. Normal casual talks, on the other hand, are meant to be unplanned and spontaneous, but one or more participants may rehearse for presentation in a discourse.

3.3. Representation

In order to achieve accurate representation of a language, a generic corpus should, in theory, include samples from a wide range of literature. Furthermore, to represent the largest amount of linguistic traits prevalent in a language, the corpus should be well-balanced, including text samples from a variety of disciplines and subject areas. Because future linguistic analysis and investigation systems based on databases will require verification and authentication of information drawn from a corpus reflecting the language in issue, the databases contained in a corpus should be authentic in their representation of the source text.

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.

4. OVERVIEW OF CONTENTS

The goal of this book is to encourage, stimulate, and challenge the reader to investigate the richness and diversity of ideas and practical applications that have been developed within the field of Corpus Linguistics, which is progressively finding its voice. This collection contains 119 articles and book chapters grouped into twelve sections that span six volumes and 2300 pages, demonstrating the amount of effort put into the field in recent years. The papers themselves were frequently difficult to locate because they had been printed in one-off events, such as conference proceedings, by the host academic institution, and many academic publishers were unable to keep up with the published works, which swiftly went out of print.

Because some things are multi-authored, there are around 125 authors represented, and 18 authors have contributed to more than one item (partly for the same reason). Despite the fact that the authors are from all over the world, the publishers are concentrated in a smaller number of locales. Some readers may be startled by the lack of American authors, but this just reflects the mostly negative attitude toward Corpus Linguistics in a tradition dominated by generative and cognitive linguists. We were questioned about the inclusion of so many papers from edited collections rather than peer-reviewed academic publications at one point during the process. The reason for this was the lack of academic journals in the field until recently (apart from the Inter-national Journal of Corpus Linguistics), as well as the wide variety of fields to which such journals belonged (studies of Text, Discourse, Pragmatics, Applied Linguistics, TESOL, ESP, Second Language Studies, and soon), and the papers often only involved corpus techniques at a very superficial level.

4.1 Language teaching

This section looks at how corpora can be used to decide what should be taught, to investigate language learners' output, and as a direct replacement or complement to standard approaches and resources in the classroom. Data-driven language learning utilizing discovery processes is described by Johns. He contends that a rule-based system (which attempts to encompass "competence") is insufficient, and that a data-driven method (accessing "performance") is more appropriate, citing the failure of co-ventures between pedagogy and Artificial Intelligence. This strategy is engaging for all levels of students and supports autonomy by stimulating students' questions, giving them with relevant material, and allowing them to utilize their wits to find their own answers. In EFL grammar teaching, Meunier assesses the pedagogical usefulness of native and learner corpora. She presents instances of corpus use in three areas of pedagogic application: curriculum design, reference tools, and classroom teaching, starting with a SLA perspective of grammar and the impact of corpus research on grammatical description. She attributes the absence of corpus use to a lack of knowledge, diminished attention to form, and the technology's unavailability, but she also cautions of corpus work's limitations, such as restricted context in concordances, which restricts understanding of text-level aspects of language. Granger begins by reviewing learner corpus research in SLA and ELT, then goes on to address corpus design requirements and analytical approaches, as well as comparing native and learner data and learner data from various student types. She considers methods of annotation (POS-tagging, error-tagging), research in pedagogy, curriculum and materials design, and classroom practices, as well as impacts on learner dictionaries, CALL programs, and web-based education, in addition to software improvements. She supports more integrated SLA, ELT, and NLP research, large corpora dissemination, in-house corpora, (automated) annotation, longitudinal studies, qualitative process-oriented studies (to augment quantitative product-oriented studies), and corpus diversification. Poos and Simpson use a corpus of academic spoken English to compare hedging across academic disciplines, and they relate to research on hedging and gender, as well as hedging in written academic speech. Gender isn't seen to play a big role in academic discourse, however there are differences in frequency and kind between physical sciences (less hedging) and humanities (more hedging). Because of their interactional and social functions, such variances are vital for EAP education. However,

REFERENCES

[1] Greenbaum, S. (1991). The development of the international corpus of English. London: Longman.

[2] Scott, M. (2012). Word Smith Tools. Version 6.0. UK.

[3] 128 Blanka Frydrychova Klimova / Procedia - Social and Behavioral Sciences 141 (2014) 124 – 128.

[4] Text STAT. (2012). Retrieved March 11, 2013, from http://neon.niederlandistik.fu-berlin.de/en/textstat/.

[5] Wu. M. H. (1992). Towards a contextual lexico-grammar: an application of concordance analysis in EST teaching.

[6] RELC Journal, 23(2), 18-34.