AUDIO CORPORA

Abstract

Corpus linguistics plays an important role in the field of language teaching, learning, and research. Corpus is a rich resource because it contains trove of authentic materials that help in exploration of the dynamic facets of the language. The use of text corpus is gained considerable traction in the field of language pedagogy and research. But audio corpora have yet to receive widespread recognition and adoption in both research and language education despite of their huge potential. This article intends to provide insights into audio corpora including their components, prominent audio corpora websites, as well as advantages, and limitations.

Keywords: Corpora, written corpora, audio corpora

Author

Dr. Aswini P
Assistant Professor
Department of Languages
Presidency University

I. INTRODUCTION

Audio corpus materials can be an invaluable source in providing authentic audio samples for language teaching and learning. These authentic audio corpus materials do not only expose learners to real world communication situations, accents, pronunciation, and cultural nuances but also bridge the gap between the classroom and the outside world. For immersing students in the authentic context, audio corpora are potential which help learners learn the language naturally as the mother tongue by exposing them to the intricacies of the authentic language. However, the audio corpora materials are not much used in the academic context compared to written corpora materials. In second language teaching context, various acoustic files are used. But these materials are far away from the real life language and they are carefully designed. These contrived materials will not prepare the learners to encounter the real life situations.

II. CORPUS

Corpus is a collection of authentic materials that are produced by the native speakers for their purpose of communication. These materials are used for linguistics analyses and for the purpose of language teaching and learning.
Nunan (1998) defines authentic materials as "which have been produced for purposes other than to teach language".

According to Carter & Nunan (2001: 68) authentic materials are "ordinary texts not produced specifically for language teaching purposes."
There are two types of corpus materials: text or transcription of audio files and audio corpus. The plural form of corpus is corpora.

III. WRITTEN CORPUS

A written corpus is a collection of texts or written documents that are used for linguistic investigations, research, and study. These texts are carefully selected and stored to represent a particular language, time period, genre, or subject matter. The written corpora are potential for linguists, researchers and language enthusiasts because they offer a plethora and diverse set of samples for analysing language patterns, usage, and various linguistic phenomena.

For written corpus, samples are drawn from various sources such as magazines, novels, book reviews, newspaper, comic books, menus, posters, postcards, conversations between native speakers, train schedules, nutrition labels, and children's books, etc.

IV. AUDIO CORPORA

Audio corpus is a collection of spoken language recordings and these audio files are transcribed and annotated for linguistic analysis and research. To study various aspects of spoken language, including pronunciation, intonation, syntax, semantics, and discourse patterns these audio recordings are crucial. These corpora provide insight into better understanding of how language is used in natural spoken communication. These audio corpora are unique and recommended by veteran professors because these materials are made

available after a thorough scanning to filter the audio files which contain foul languages. It is also important to say that during this process, the conversations are neither changed nor cropped. As experts say that if any modification is carried out, the authenticity will be lost and it will not be considered as authentic materials. These scanned corpora materials are stored on the computer for the purpose of research and teaching and learning. Exposing students to these materials for an extensive listening would help them grasp the dynamic nature of the language such as accents, intonation, pauses, and other nuances that are absent in the written communication.

Audio corpus materials are drawn from various sources. The common sources from which audio files are collected interviews, lectures, speeches, conversations, radio broad casts, news reports, podcasts, audio books, online forums, chat logs, social media platforms, voice assistants like Amazon Alexa, and Apple Siri, etc.

V. COMPONENTS OF AUDIO CORPUS

1. **Audio Recordings:** The foundation for audio corpus is audio recordings. These audio recordings span from short phrases to lengthy conversations, depending on the research goals.

2. **Transcriptions:** Transcriptions convert spoken language or audio content into written text. These are all textual representation of the audio content which helps researchers to analyze and study linguistic features. Further, these materials are immensely useful in language learning when learners have difficulty to comprehend native speakers' language.

3. **Annotations:** Annotations provide additional information about the audio files. The information can include speaker's identity, intonation patterns, pauses, grammatical structures, and more.

4. **Metadata:** Metadata offers contextual detail about recordings, such as location, date, time, participants' demographic details, and the communicative context.

VI. WHERE CORPUS MATERIALS CAN BE APPLIED?

Corpus materials have an extensive range of applications. Audio corpus materials are used in the following areas:

1. **Linguistic Research:** Audi corpora are essential for studying the sounds of speech. Linguists use audio corpora to analyze pronunciation variations, speech patterns, intonation, and prosody. They also investigate how phonetic features change in different linguistic contexts or among different speakers.

2. **Psycholinguistics:** Audio corpus is paramount importance in psycholinguistics field as it aids in the study of language processing, acquisition, production, and variation. By analyzing the samples of audio files, researchers can examine how people comprehend, produce, and acquire language.

3. **Language Teaching:** Audio corpus materials can be effectively used in language teaching and learning. For example, authentic audio samples for listening comprehension, pronunciation practice, vocabulary acquisition, and cultural understanding.

4. **Speech Technology:** Audio corpus is used to train, test, and develop speech technology systems. It provides diverse spoken data for training speech recognition models, building language models, identifying speakers, recognizing accents, analyzing emotions, and more.

5. **Cultural and Societal Studies:** Audio corpus materials provide insight into other cultures such as cultural norms, social dynamics, and identity through language use.

VII. WELL KNOWN AUDIO CORPORA WEBSITES

Here are a few websites that provide audio corpora for various speech and language processing tasks. There are different corpora websites which have differences in their coverage. The following ones are the prominent websites of audio corpora:

1. **Longman Spoken American Corpus:** This corpus is owned by Pearson Education and audio files were gathered by Professor Jack Du Bois and his team at the University of California, Santa Barbara. But this corpus is not easily accessible. This corpus contains collection of everyday conversations of more than 1000 Americans of various age groups, with different levels of education, and ethnicity. The conversations run about four hours at a time and these audios are recorded as obtrusively as possible.

2. **British National Corpus (BNC):** This website offers a collection of audio recordings sourced from the spoken portion of the British National Corpus. These recordings are digitized from analog audio cassette tapes stored at the British Library Sound Archive. In addition, the site provides related transcription and annotation files that are generated as part of the Mining a year of speech project. This website contains two types of audio files: “context- governed” recordings and “Demographic” recordings.

3. **Michigan Corpus of Academic Spoken English (MICASE):** MICASE is a preset-day corpus which contains recordings that span about 190 hours. The audio files are recorded within the premises of the University of Michigan. The audio files contain conversations and lectures of native and a few non-native speakers including faculty, staff and students of all levels. This corpus contains 15 varied speaking scenarios from across the major academic domains such as medicine, business, law, and dentistry. This corpus is available with free of cost.

4. **Santa Barbara Corpus:** Santa Barbara Corpus of Spoken American English (SBCSAE) is distinguished from other corpora because it contains only audio files. These audio files are gathered and electronically stored by professors at California University. This corpus website designed in such a way to provide a hassle free access to the contents. The electronically stacked audio files are drawn from heterogeneous background such as from myriad regions, gender, occupations, ages, and social backgrounds. This corpus is freely available and users have direct access to a considerable amount of audio materials.

Acoustic models available with transcripts, and provides a brief introduction to the contents of audio files.

5. **The Nationwide speech Project (NSP):** It is a gamut of spoken language which represents 6 regional varieties English across United States. The speech material encompasses isolated words, sentences, passages, and interviews. The corpus was created for the purpose of utilizing it in acoustic and perceptual investigations concerning regional dialects differences within the United States.

VIII. ADVANTAGES OF AUDIO CORPUS MATERIALS

One of the main benefits of using authentic materials in the classroom is to *'expose'* the learners to authentic language use and help them communicate effectively in real life situations. According to Su (2008) the main advantages of using authentic materials in the classroom are as follows:

Authentic Materials:

- Expose students to real life conversations that are related to learners' every day needs. Learners get exposure to the real word intercultural discourse (Kilickaya, 2004; Martinez, 2002; Morrision, 1998; Peacock, 1997)
- Serve as a means of practicing small- scale abilities like scanning or micro skills associated to listening such as listening to news reports, identifying the names of people or countries (Martinez, 2002; Peacock, 1997).
- includes diverse range of textual genres and linguistic approaches that are not easily available in the conventional educational resources (Martinez, 2002; Peacock, 1997; Grundy, 1993; Sanderson, 1999)
- Contain interesting topics and encourage reading for pleasure and they provide valid linguistic data (Dumitrescu, 2000; Martinez, 2002; Peacock, 1997).
- Have an intrinsic value and keep learners updated about what is happening around the world (Martinez, 2002; Peacock, 1997; Sanderson, 1999) .
- Offer an opportunity to distribute information and cross- cultural understanding (Gebhard, 1996).
- provide cultural information from real context and open the door for cultural adaptation, language comprehension, and language use (Sanderson, 1999; Grundy, 1993; Duquette, et, al, 1987).
- Can be employed for a specially designed curriculum and act as a conduit connecting the classroom with the real world scenarios (Peacock, 1997).

In the language classroom authentic materials can be used to design various activities in order to enhance listening and speaking. In other words, these materials are flexible, can be adapted according to learning objectives, and are not bound by the limitations of the textbook format.

IX. ISSUES IN USING AUDIO CORPUS MATERIALS

While authentic materials are fruitful in the realm of language teaching, it has certain limitation associated with their utilization in the classroom.

According to Martinez (2002) authentic materials are culturally biased and it will be challenging for learners to comprehend in the classroom settings. Moreover, if these materials are not presented and explained within their genuine cultural context, they can potentially convey wrong notions about other cultures.

Designing activities based on authentic materials demands a significant amount of time from the teacher. In Miller's(2005) words, authentic materials can be deemed as "excessively challenging and time- intensive to choose, modify, and ready for use."

As stated by Richards (2000). Another downside of authentic materials is that these materials contain intricate and opaque language, as well as less used vocabulary items and complex language structures. These elements can pose challenges for learners at primary level with reduced enthusiasm and de-motivation to engage in active learning.

Despite the above mentioned challenges associated with using authentic materials in language learning, they can be extremely valuable resources since they provide numerous advantages.

X. CONCLUSION

Audio corpora present a rich and authentic resource in the domain of language education and linguistic research. The advancement of technology has democratized access to audio corpus poised a significant growth. Researchers and educators can harness audio corpora to gain deeper insight into language variation, evolution, and usage. Language teachers can utilize these materials in order to give novel experience to the learners as well as to learn the language naturally. While audio corpora provide numerous advantages, it is essential to keep in mind the limitations they pose such as comprehension of native speakers' language, identifying appropriate audio files to design activities, and sustaining learner's listening engagement over an extended time, etc. By understanding and addressing the difficulties associated with audio corpora, we can fully leverage their applications and benefits in the modern world.

REFERENCES

[1] Dumitrescu, V. "Authentic Materials: Selection and Implementation in Exercise Language Training." Forum, vol. 38, no. 2, 2000.
[2] Grundy, P. Newspapers. Oxford University Press, 1993.
[3] Gebhard, J. G. *Teaching English as a Foreign Language: A Teacher Self-Development and Methodology Guide*. The University of Michigan Press, 1996.
[4] Kilickaya, F. "Authentic Materials and Cultural Content in EFL Classrooms." *The Internet TESL Journal*, vol. 10, no. 7, 2004, pp. 1-6.
[5] Lindquist, H. *Corpus Linguistics and the Description of English*. Edinburgh University Press, 2009.
[6] Martinez, A. G. "Authentic Materials: An Overview." *Karen's Linguistic Issues*, 2002.
[7] Morrison, B. "Using News Broadcasts for Authentic Listening Comprehension." *English Language Teaching Journal*, vol. 43, no. 1, 1989, pp. 217-221.
[8] Miller, M. *"Improving Aural Comprehension Skills in EFL, Using Authentic Materials: An Experiment with University Students in Nigata, Japan."* Master's thesis, University of Surrey, 2005.
[9] Nunan, D. *Designing Tasks for the Communicative Classroom*. Cambridge University Press, 1989.
[10] Peacock, M. "The Effect of Authentic Materials on the Motivation of EFL Learners." *ELT Journal*, vol. 51, no. 2, 1997, pp. 144-156.

[11] Richards, C. "Hypermedia, Internet Communication, and the Challenge of Redefining Literacy in the Electronic Age." *Language Learning & Technology*, vol. 4, no. 2, 2000, pp. 59-77.
[12] Sanderson, S. *Using Newspapers in the Classroom*. Cambridge University Press, 1999.
[13] Su, Sh. "Attitudes of Students and Instructors toward Authentic Materials in Selected Adult TESL Programs." 2008. Retrieved from http://ir.lib.au.edu.tw/bitstream/987654321/2628/1/

