Providing a Suitable Method for Allophonic Labeling of Speech Corpuses According to the IPA System

Tahere Ahmadi[1]
Batool Alinezhad[2]
Hossein Karshenas[3]
Bagher Babaali[4]

Received:11/09/2019
Accepted:29 /01/2020
Article Type: Research

Abstract

The corpus is a collection of spoken and / or written texts that can be used for linguistic analysis. More precisely, it can be said that these texts are purposefully labeled and categorized based on specific rules and allow the user to do various studies. Corpus linguistics is a branch of applied linguistics that examines and compares different aspects of linguistic data, and, of course, corpora are integral tools of this branch of linguistics. Due to the increasing role and importance of corpus linguistics in development of various sciences in recent decades, the produce and development of various linguistic corpora has been one of the priorities of scientists and researchers in different languages during these years.

After the creation of speech processing systems since about two decades ago, the use of context-dependent methods has become particularly prominent in an effort to increase the accuracy of these systems and some special studies conduct in linguistics,. One of the best ways to achieve this, is to use corpora that, have special labels in addition to segmentation at the phoneme level, to indicate the differentiation of various allophones. These allophnescan only be achieved by obtaining the necessary phonological rules. In linguistics, this process can be called allophonic labeling of corpus.

About 10 years after the introduction of allophonic corpora in the world, no allophonic labeling has been performed for any of Persian language corpora yet. The small Farsdat corpus is the main spoken corpus in Persian. Hence, the need to equip

[1] MA, Computational linguistics, Department of Linguistics, Faculty of Foreign Languages, University of Isfahan,Isfahan, Iran; pazhvak_ta@fgn.ui.ac.ir

[2] PHD, Linguistics, Associate Professor and Faculty Member in Department of Linguistics, Faculty of Foreign Languages, University of Isfahan, Isfahan, Iran, (corresponding author); b.alinezhad@fgn.ui.ac.ir

[3] PHD, Artificial intelligence, Assistant Professor and Faculty Member in Department of Artificial intelligence, Faculty of Computer, University of Isfahan, Isfahan,Iran; h.karshenas@eng.ui.ac.ir

[4] PHD, Artificial intelligence, Assistant Professor and Faculty Member, Faculty of Mathematics, Statistics and Computer Science, University of Tehran, Tehran, Iran; babaali@ut.ac.ir

this corpus with allophonic labels to increase the accuracy, to improve the performance of speech processing systems , and to produce specific study, research programs, and tools in linguistic is obvious. In order to elucidate the method proposed in the present study for allophonic labeling of phonemic corpuses, and in parallel for equipping the Persian language with at least one allophonic corpus, the steps of the task are precisely performed on the small Farsdat phonemic corpus. The corpus is one of Persian-language corpora in the last two decades that consists of 6080 sentences spoken by 304 Persian speakers. The speakers of this corpus have indeed one of the most widely spoken dialects in Persian and all of sentences in this corpus, are segmented in to different levels. The segmentation of sentences in word and phoneme levels results in their efficiency in various speech processing systems, such as speech recognition systems, broad transcription systems, and text-to-speech systems. Moreover, the small Farsdat corpus has the potential to be used in the systems.

The suggested solution to prepare an allophonic corpus is to implement a program using the rule-based method and applying it on the phonemic corpus to add allophonic labels on it. The basis of the rule-based method in this research is access to rules for converting phonemes into allophones. After compiling these rules from the resources available in each language and preparing the appropriate settings (for implementation), the program is implemented. Finally by applying this program to the phonemic corpus, an allophonic corpus is prepared.

As noted, special phonological rules are required to convert phonemes into allophones in Persian and to add allophonic labels to the small Farsdat corpus. The purpose of this research is not to study phonemes based on acoustic and laboaratory approaches in order to obtain Persian allophones; but rather to formulate and synchronize phonemes identified in various studies and then to adapt them to the International Phonetic Alphabet System. This ultimately leads to provide a standard set of allophones as far as possible and to achieve the phonological rules necessary for converting phonemes into allophones in Persian (based on existing studies.

Although one of the limitations of this study is its incompleteness regarding the extraction of different allophones in Persian, the implemented program has the capability to be updated. if any studies are carried out in the field of allophones to supplement the existing theoretical resources in the future, it has the possibility to be to modified or to be enhanced regarding the performance . The present study may also highlight the need for more recent linguistic experiments and the use of more accurate tools and facilities to identify Persian phonemes. This can increase the motivation of phonetics and phonology researchers to take more practical steps in this field as well.

After providing the necessary preparations in the phonemic corpus (such as the syllable segmentation) and implementing the above rules, the allophonic labels can be added to the phonemic corpus by implementing this program on it.

Keywords: Phoneme, Allophone, Corpus, IPA system