Building a Phonetic Corpus from Librivox Audiobooks

"Librivox.org" is a non-profit library of free audiobooks recorded by volunteers. The website was initiated in 2005 by Hugh McGuire. It offers more than 12 000 finished projects, most of which are novels in English.

The current project aims to explore the possibilities of creating a specialized corpus on the basis of the aforementioned audiobooks library. The corpus would be especially useful for phoneticians and other linguists, but could be also applied in other domains related to natural language.

At the initial stage of the project, 105 English audiobooks were downloaded from 'librivox.org'. In order to obtain a sample representing different dialects and genders, only these audiobooks were chosen which were read by groups of readers, rather than an individual person. After that, the corresponding text versions of the novels were found at 'gutenberg.org'. With the use of a Python script, the texts were divided into syntactically and prosodically independent units. Next, these units were automatically aligned with the corresponding parts in audiobooks using *Aeneas,* which is a Python/C library designed to automatically synchronize audio and text. Additionally, with the use of various scripts written in Python, all the text units and the corresponding recordings were classified according to numerous criteria, such as context (narrator vs. dialogue), pragmatic function (statement/directive vs. quesiton vs. exclamative statement), reader's gender (female vs. male), reader's dialect (American vs. British vs. Australian vs. non-native), author's gender (female vs. male), number of words, syllables and phonemes, duration, etc.

On the basis of the database obtained, it is possible to create a free online corpus offering versatile functionality. A prototype version of such a tool is currently under development. It aims to enable searching for individual words and phases in the audiobooks. The fragments containing such words or phrases will be available for audio playback directly from a web browser and for download as mp3 files. Moreover, search options will allow filtering the results according to many criteria, such as the ones mentioned previously.

Possible future developments include enlarging the database to about 1000 audiobooks, which would be 10 times more than in the current version. The predicted duration of audio materials for such a sample would be above 10 000 hours, and the number of words in the corresponding texts would be above 100 000 000. To the best of the author's knowledge, this would result in the largest phonetic corpus for the English language currently available. The tool would be comparable in volume to large text-based corpora, such as British National Corpus. Moreover, text alignment could additionally be performed on the word level and part-of-speech tagging could be introduced. This would allow the user to search for individual words and phrases in a much more flexible way.