Data paper

Title: Comparing Queen Elizabeth II Received Pronunciation Features (1950s, 1990s, 2010s)

Bibi Malaika 11542742, Gurung Aakriti 11377977, Rai Ninam 11455884, Valleramos Janelle Joyce 11455107, Zaheer Aqsa 11455638

[a]*Faculty of Humanities, The Education University of Hong Kong, Tai Po, Hong Kong*

*Corresponding author: first name, initials, surname; email address

Aakrtiti Gurung (s1137797@s.eduhk.hk)

Aqsa Zaheer (s1145563@s.eduhk.hk)

Joyce Janelle Valleramos (s1145510@s.eduhk.hk)

Malaika Bibi (s1154274@s.eduhk.hk)

Ninam Rai (s1145588@s.eduhk.hk)

Author roles

Please list the roles for each author

Aakriti – Methodology, Investigation

Aqsa – Abstract, Overview (context)

Joyce – Methodology, Investigation

Malaika – Reuse Potential

Ninam – Methodology, Investigation, Data Curation

Word Count: 1063

Abstract

This dataset offers a comprehensive analysis of Queen Elizabeth II's Received Pronunciation (RP) features across three distinct decades: the 1950s, 1990s, and 2010s. The data was meticulously gathered from YouTube, focusing on her speeches or messages during these periods, and processed for audio and text analysis. Stored in a GitHub repository, the dataset includes raw and processed audio files, transcriptions, and annotated analyses in JSON format. Its reuse potential spans linguistic research, educational purposes, and offering insights into phonetic shifts over time.

Table of Contents

(1) Overview

Repository location *– Github*

Context

This dataset was produced as part of a research project aimed at examining the evolution of Received Pronunciation in Queen Elizabeth II's speeches over several decades. At least twenty minutes of video was collected from each of three time periods, resulting in a total of 1 hour of recorded material (data).The project focuses on linguistic shifts and phonetic features, contributing to a broader understanding of language change and sociophonetics. The dataset has potential applications in academic research, linguistic studies, and historical analyses.

(2) Method

Steps **–**

Video Acquistion:

First, we first look for Queen Elizabeth II speeches videos in YouTube. After finding appropriate vidoes of different time period (1950s, 1990s, 2010s) worth one hour long total. All the vidoes were downloaded using ‘yt_dlp’ library in Google Colab. It is stored into a shared Google Drive folder and then uploaded to GitHub.

Audio Extraction & Transcription:

Next, to extract the audio tracks, ‘moviepy’ library was used. Finally, OpenAI whisper library was employed to transcribe text from audio. The transcription is edited to ensure

that it is only the Queen's words that can be read. Our extracted audio and transcription files are also stored into the Google Drive folder and later uploaded into GitHub.

Selection of words:

After these first three initial steps, specific words were chosen from each time periods to identify the phonetic features of RP and measure if there is any shift in Queen Elizabeth II phonetic features as time passes. These words were chosen as they are known to show variation in Received Pronunciation (RP) based on our research (Payne & Laura, 2023).

Audio Conversion & Audio Segmentation:

Full audio recordings mp4 (raw) wre first converted to wav files using 'cloudconvert' and saved as (QueenYear_VideoName = Queen1953_ChristmasMessage). Then, selected words were further extracted from the full audio recordings using Praat and saved as individual audio.wav files e.g., (QueenYear_Videoname(word) = Queen1953_Christmas (world)). These files can be found in our Google Drive, with links accessible through our GitHub.

Data Annotation:

The transcript in text file format was processed into JSON format and annotated to provide a detailed and organized analysis of Queen Elizabeth II's speeches. The metadata section in annotation offers important context, such as the title, date, occasion, and speaker details, ensuring users has some context in the speech's background. Each paragraph is broken down and an analysis of its purpose, tone, or theme is made, allowing for precise identification of recurring ideas. This enables users to search or filter by specific elements, such as joyful tones or themes like progress. The summarized analysis highlights overarching themes (e.g., unity, service), rhetorical

techniques (e.g., inclusive language), and the speech's relevance according to the year. The JSON format was chosen for its machine-readability, scalability, and clarity, making it useful for education, research, or AI training. The words has been transcribed in Peter Roach's International Phonetic Alphabet to identify the features of RP.

Sampling strategy – All our videos were selected to preserve and compare the RP features of Queen Elizabeth II during these three distinct time periods 1950s, 1990s, and 2010s. Most videos were chosen based on her speeches and messaged and clarity of the audio. From all the three chosen time periods, we have a total of one hour worth of data. For her 1950s we have chosen 3 speeches: Coronation speech and Christmas messages of 1953 and 1957. For her 1990s speeches, we have chosen Annus Horriblis, Congress and BBC news speech. Her 2010s speeches, we picked Christmas message of 2011 and 2013, United Nations speech, 400th anniversary speech and Summons.

Quality control – The audio was all processed into WAV format for easy data collection/storing in Github. The context of the Queen's speech was mostly either the Christmas Speech or official addresses, to ensure some amount of variety while keeping the data consistant. The choosen time period (1950s, 1990s, 2010s) has a balanced time coverage to ensure an even distrubtion of samples. The transcript ensures the text is only the speech spoken by the Queen.

(3) Dataset Description

Repository name – Github

Object name – https://github.com/gurungaakriti2001/GodSaveTheQueen

Format names and versions –

Videos (Raw): mp4

Audio (Raw): WAV

Audio (Processed – Selected Words): Prat (extract the selected words), WAV

Transcriptions (Raw): sty.md

Annotations (processed): JSON

Formant Data: Praat tables

Creation dates – [The start date: 2025-03-11]] [The end date: 2025-04-18]

Dataset creators:

Joyce - Data Collection (1950s Queen Elizabeth II Speeches), Audio Extraction, Transcription

Ninam - Data Collection (1990s Queen Elizabeth II Speeches), Audio Extraction, Transcription, Text Annotation

Aakriti – Data Collection (2010s Queen Elizabeth II Speeches), Audio Extraction, Transcription

Language – For our dataset, English is used. The contents of our INTRODUCTION.md file is in English and the transcription of her speeches, both raw and processed, are kept in English to establish authencity when analyzing. All the folders in our dataset are named "Queen Elizabeth II Speeches [Raw/Processed] [File Type] [Year]". Aside from our folder names, we have also named our files in English: "Queen[year]_[speech type]". This is especially prominent in our folder of our processed audio files, in which we included our chosen words from her speeches into the file names (e.g. Queen[year]_[speech type](*word*)). As her choice of language in her speeches was English, it is to ensure easier navigation when analyzing as well as comparing her pronunciations of the same words through different periods of her reign.

License – MIT License

Publication date – 2025-04-19

(4) Reuse Potential – The existing dataset contains various reusability possibilities for historical linguistic analysis and educational use together with AI training applications:

1. Through studying linguistic modifications and phonetic developments within RP throughout multiple time periods linguists gain substantial benefits from their audio analyses and annotation work and measurement data.
2. Historians and sociolinguistics can use the current research data to investigate relationships between generalized social-economic patterns and linguistic developments including cultural evolution and global trends.
3. Teachers can use the data to improve student understanding of phonetics along with sociolinguistics while teaching British cultural components.
4. The JSON-formatted annotation in transcripts allows professional linguistic experts and AI processing systems to complete targeted analyses through simplified data processing steps.

   Lastly, this data faces technical barriers because its restricted sample selection and analysis parameters limit it to traditional RP historical research only.

Acknowledgements

We acknowledge that we used AI https://poe.com for annonation (JSON), https://openai.com for extracting text from audio, and Praat for acoustic analsyis.

Competing interests

The authors have no competing interests to declare.

References –

Payne, & Laura. (2023, June 23). *Received Pronunciation (RP) | Accent, audio Examples, IPA, & Definition*. Encyclopedia Britannica. https://www.britannica.com/topic/Received-Pronunciation