Overview
• Incorrect reading with a screen reader confuses visually impaired persons and children with reading difficulties in understanding written text
• We have built two large-scale pronunciation annotated corpora in Japanese: Book Title Corpus (336M char) & Aozora Bunko Corpus (52M char.)
• We used the obtained corpus to train and evaluate BERT-based pronunciation classifiers, and obtained macro accuracy of 0.939

Research Background
Screen readers are essential for visually impaired persons to read text via speech
• [Japan] a law on act to further the improvement of reading environments for visually impaired persons
Challenge: heteronomous logograms (kanji) causes serious troubles in reading text by screen readers

Method to obtain corpora with word-level pronunciation annotations
We exploit existing resources with sentence and document-level pronunciation annotations to obtain word-level annotations
• Book titles compiled by National Diet Library (NDL), which cover all books and magazines since the late modern era (sentence-level annotations)
• Fiction and non-fiction books in Aozora Bunko (Japanese digital library) and its Braille translations (document-level annotations)

Method on predicting pronunciation using BERT
We evaluate the utility of our corpora on pronunciation prediction
• Train/dev/test data: 456,223, 152,095, and 152,180 sentences
• Model: Pretrained BERT [https://huggingface.co/cf-tohoku] for sequence labeling
• Target: 93 heteronyms in the subword vocabulary of the pretrained BERT

Analysis on pronunciation distributions
We see pronunciation distributions in the obtained corpora
• Target: 203 heteronomous logograms (kanji) extracted from applied rules for characters and “Yomi” [National Diet Library, 2021]
• Exclude compound nouns such as (国立駅, Kunitachi eki), since their pronunciations can be unique (if a dictionary covers)

Experiments on predicting pronunciation using BERT
We evaluate the utility of our corpora on pronunciation prediction
• Train/dev/test data: 456,223, 152,095, and 152,180 sentences
• Model: Pretrained BERT [https://huggingface.co/cf-tohoku] for sequence labeling
• Target: 93 heteronyms in the subword vocabulary of the pretrained BERT