Cultural Analytics Dataverse 13 October 2021

Replication Data for: "Audiobook Stylistics: Comparing Print and Audio in the Bestselling Segment" (*Journal of Cultural Analytics*, 2021)

Karl Berglund and Mats Dahllöf
Department of Literature and Department of Linguistics and Philology
Uppsala University

This repository contains code and frequency data that can be used to reproduce results and figures published in the paper "Audiobook stylistics: Comparing print and audio in the bestselling segment". The frequency data are derived from the corpus of recent bestselling fiction. That corpus is protected by copyright and cannot be shared, but the code for creating the frequency data is supplied.

PLEASE NOTE CONTENT WARNING: While use of offensive and harmful language has not been relevant for the discussion in the article, we present the frequency data for replication purposes "as-is" based on the actual text of the works of fiction included in the corpus.

The corpus exists in the form of text files extracted from .epub files. The works were tagged using the "Swedish annotation pipeline". See R. Östling (2018), "Part of Speech Tagging: Shallow or Deep Learning?," *North European Journal of Language Technology* 5. (`https://github.com/robertostling/efselab`)

The tagging generated files in the so-called conll format. This is an example excerpt, words 11 to 17 of a sentence.

```
11 bara bara ADV AB _ 10 advmod _ _
12 på på ADP PP _ 14 mark _ _
13 att att PART IE _ 14 mark _ _
14 få få VERB VB|INF|AKT VerbForm=Inf|Voice=Act 17 aux _ _
15 se se VERB VB|INF|AKT VerbForm=Inf|Voice=Act 17 aux _ _
16 den den PRON PN|UTR|SIN|DEF|SUB/OBJ Definite=Def|Gender=Com|Number=Sing 17 nsubj _ _
17 släppa släppa VERB VB|INF|AKT VerbForm=Inf|Voice=Act 10 xcomp _ _
```

The replication code (Python), metadata, and frequency data are bundled in the `audiobstyle.zip` file.

The Python code is located in `./src/`.

Metadata are located in `./meta/`. The file `book_data_211008.csv` provides basic metadata about the works. Its fieldnamed should be easy to comprehend. List positions are recorded in `list_positions_201203.csv`, with the columns `hb_year`, `hb_pos`, `pb_year`, `pb_pos`, `str_year`, `str_pos`, for the year of its best performance and position, for the hardback, paperback, and streaming toplists, respectively.

Frequency data generated from the corpus is located in `./dicts/`. The code used for generating it from the conll files is supplied. Executing the file `generate_frequency_data.py` performs this step, when the corpus is available.

The results reported in the article can be generated from the data in `./dicts/` and `./meta/` by running `main.py`. The results will then be stored in the directory `./results/`. The .zip file provided here includes the result files, as described below.

`./results/cles_tables.tex` is the LaTeX code for the for CLES tables generated. Some of these tables appear in the article. Their interpretation is explained there.

`cles_tables.pdf` is the pdf generated by means of MikTeX/XeLaTex from `./results/cles_tables.tex`.

`./results/all_cles.csv` contains CLES scores for all measures and comparisons. CLES scores expressed in permille.

`./results/plots/` contains a number of plots, three of which appear as figures in the article.

`./results/rankings_by_feat/` contains for each measure, a list (csv) of the works ranked by that measure.