Supplementary Materials

for "Quotation and Narration in Contemporary Popular Fiction in Swedish – Stylometric Explorations" (DHNB 2022 submission)

Mats Dahllöf, Uppsala University, Sweden

This repository of Supplementary Materials provides details on the corpus and a more complete presentation of the results in the article.

1. Metadata

`dhnbcorp.csv`

A csv listing of the works in the corpus along with some metadata. The following keys are used:

id project-internal identifier.

isbn "S" before the number means that the entry is a series of episodes concatenated into a season document. The number is the actually the isbn of the first episode. For isbns for episodes, see below.

title

all_authors

publisher

publication_date

author_gender

format bestseller, beststreamer, or (Storytel) original.

category CRIME, PRESTIGE, OTHER

orthography quot(ation-based), dash(-based), or other

`episodes_seasons.csv`

This file provides information about the episodes making up works published as a series of episodes. We concatenated all the episodes of each season into one document in this study.

`corpfeatvals.csv`

All the features, expressed as parts per million, for each book, with some metadata overlapping with `dhnbcorp.csv`. Feature names as explained in the article.

2. Pipeline for classification of quoted inset and narrative frame material

The pipeline for classification of quoted inset and narrative frame material is trained on the corpus described in the main article. We cannot share this data due to copyright reasons.

The first step in the pipeline is to download the Swedish "talbanken 1.0.0" model for the Stanza system (`stanfordnlp.github.io/stanza/`). This resource is then used to tag the texts with part-of-speech (POS) labels. The result is stored as csv files.

Secondly, the pipeline downloads the classifier for telling quoted inset and narrative frame material apart for dash-orthography works. It is applied to the tagged csvs, producing new csvs with a qi/nf column added. A rule-based module is then used to tag the quotation-mark-orthography works.

The qi/nf classification results in assignment of the following labels, which are attached on the word level.

- "`nondial`" for NF only paragraphs, i.e. those not beginning with a dash (–) or containing quotation marks (").
- "`inquote`" and "`nonquote`" for QI and NF material, respectively, in typical QI paragraphs i.e. those beginning with QI material as indicated by an intial quotation mark or dash. (The classifier targets the QI/NF separation in the dash-orthography variety of such paragraphs.)
- "`niq`" for word in other paragraphs, i.e. those with non-initial explicitly quoted text or non-matching quotation marks.

The Python code for the pipeline is available in the file `dialtag_pipeline.zip`, and is executed by running the file `main.py`. The code also involves a package called `dialtag`.

The implementation is intended for the kind of text volumes relevant for the main article. It is somewhat wasteful both as regards time and memory consumption.