UPPSALA UNIVERSITET 11 April 2019

Swedish Prose Fiction with Modern Spelling From Litteraturbanken [Data Set] (Version 1.0.0)

This dataset contains the prose texts in modern (post 1906) spelling from Litteraturbanken – "the website for reliable digital versions of Swedish classics" (litteraturbanken.se) – which can be used under the Creative Commons Attribution-NonCommercial-ShareAlike license. The 118 texts have been retrieved as fairly plain textfiles from the XML files supplied by Litteraturbanken.

Curation (of the dataset in this form): Mats Dahllöf and Karl Berglund, Uppsala University.
(mats.dahllof@lingfil.uu.se, karl.berglund@littvet.uu.se)

Files

`lbpf.zip` A zip file with the 118 text files of the corpus.

`corpusLBPF.cvs` Metadata as described below.

`docLBPF.pdf` This file.

Metadata (`corpusLBPF.cvs`)

The textfiles are identified by the internal numbering of Litteraturbanken, e.g. `lbpf11779.txt`. In some cases only a part of a work is prose. In such cases it is stored in a file like `lbpf262944_273_294.txt`, where 273–294 is the page range.

Metadata are collected in a CVS file `corpusLBPF.cvs`, with seven fields as follows (illustrated by five examples). (Blanks added for column alignment.)

```
1648810,     "Agrell, Alfhild",   K,1917,     Nordanfrån,                  N,EQ
1435563,     "Andersson, Dan",    M,1914,     Kolarhistorier,              N,EQ
81740,       "Andersson, Dan",    M,1918,     De tre hemlösa,              R,DASH
8203287,     "Strindberg, August",M,1884-1886,Samlade Verk 16. Giftas I-II,N,DASH
8203269_7_82,"Strindberg, August",M,1903,     Ensam,                       N,DASH
```

(1) Identifier: internal numbering of Litteraturbanken.

(2) Author.

(3) Author's sex: (M) male, (K) female.

(4) Year of first publication. (The text is often from a later edition.)

(5) The short form of the title (from Litteraturbanken).

(6) Genre: N for short story (collection) (Swedish "novell (samling)") or R for novel (Swedish "roman").

(7) Dialogue style: EQ indicates explicit marking of dialogue utterances with quotation marks. DASH is for books where a dash indicate the start of an utterance.

The texts

The text files contain paragraphs and non-enumerative headings. (*Tredje kapitlet*/*Third Chapter* would be counted as an enumerative heading and would not be included in the textfile.)

Quotation marks: Guillemets (used as quotation marks) are converted to ordinary Swedish style double quotation marks.

So,

```
»Inte ska Selma gråta för dä,» sade hon. »Jag ska bära'na.»
```

is rendered as:

```
”Inte ska Selma gråta för dä,” sade hon. ”Jag ska bära'na.”
```