New file format for the Onomastica lexicon

Språkbanken

September 2021

Introduction

To make the Onomastica lexicon more available to users, Språkbanken has parsed the original *.on*-files, and has generated a *.csv*-version of the lexicon, *onomastica.csv*. This document briefly describes this version of the lexicon. For more detailed information about the different data fields, see the original documentation.

Note that the quotation character is set to single quotes, *',* in the *.csv*-file, as double quotes are used to represent stress and tones in the SAMPA transcriptions.

Fields

The *.csv*-file has 556 281 data lines, one for each name in the lexicon. It has the following columns: *word_id, original_file, onomastica_id, orthographic_name, frequency, name_class, etymology, pron_0, pron_id_0, quality_level_0, transcriber_id_0, pron_1, pron_id_1, quality_level_1, transcriber_id_1, pron_2, pron_id_2, quality_level_2, transcriber_id_2, pron_3, pron_id_3, quality_level_3, transcriber_id_3*

- *word_id*: A word identifier created when generating this new version of Onomastica. It is a counter from 1 to 556281, the length of the word list.
- *original_file*: The name of the original *.on*-file
- *onomastica_id:* The *ENT* field in the original files. This is an identifier for names. Note that names spelled similarly across name classes appear to have the same *onomastica_id*
- *orthographic_name*: The *LBO* fields in the original files. The orthographic form of the name
- *frequency:* The *FQO* field in the original files. The number of occurrences in the source data.
- *name_class*: The *CT0-3* fields in the original files, classifying names into categories such as *street*, *surname* etc. We have collapsed these fields into one, since they do not seem to contain distinct values.
- *etymology*: A language code indicating the origin of the name. This field is a collapsed version of the *ET0-3* fields in the original files, which do not seem to contain distinct values.

- *pron_0*: The SAMPA transcription of the basic pronunciation of the name. Corresponds to the *NO0* field in the original files. For the vast majority of the names, this is the only transcription.
- *pron_id_0*: An identifier for *pron_0*. This identifier has been created when generating the *.csv*-file. Note that identical transcriptions have the same identifier across names, so it is possible to use this identifier to collect homophonous names.
- *quality_level_0*: An indication of the quality level of *pron_0*. This corresponds to *QU0* in the original file.
- *transcriber_id_0*: An identifier for the transcriber(s) of *pron_0*. This corresponds to *WH0* in the original file.
- *pron_1-3*: Alternative pronunciations 1-3, which correspond to *NO1-3* in the original files.
- The remaining fields correspond to *pron_1-3* in the same way as *pron_id_0, pron_id_0, quality_level_0* and *transcriber_id_0* correspond to *pron_0*.