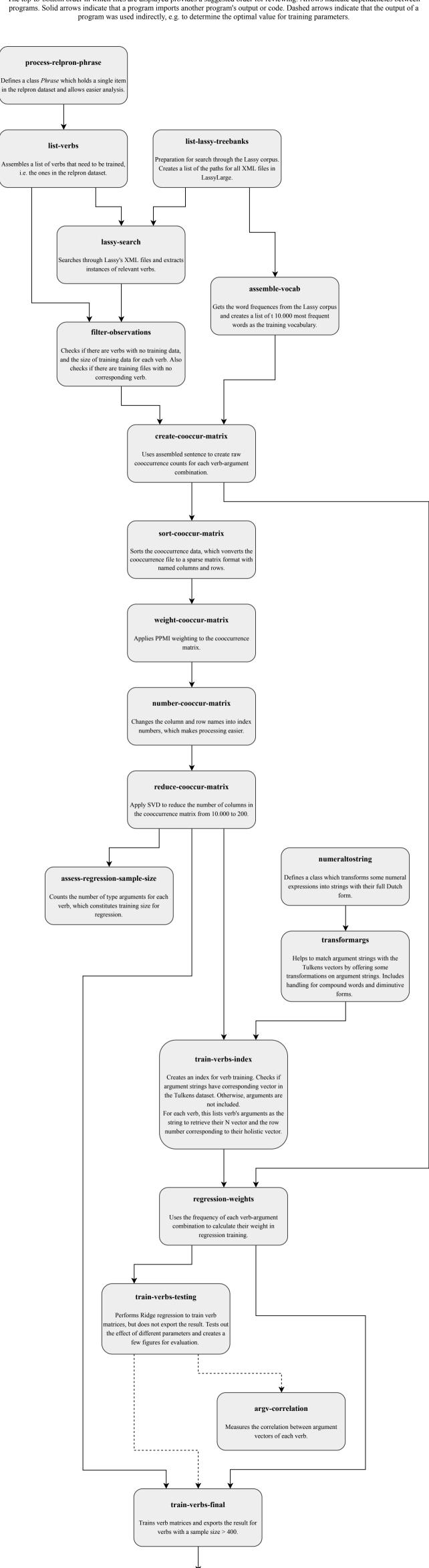# Structure of python files

The chart below illustrates the order in which python files were used in the analysis and their relation to each other. It is advised to keep this structure in mind when reviewing the code.

The top-to-bottom order in which files are displayed provides a suggested order for reviewing. Arrows indicate dependencies between programs. Solid arrows indicate that a program imports another program's output or code. Dashed arrows indicate that the output of a program was used indirectly, e.g. to determine the optimal value for training parameters.

process-relpron-phrase

Defines a class *Phrase* which holds a single item in the relpron dataset and allows easier analysis.

list-verbs

Assembles a list of verbs that need to be trained, i.e. the ones in the relpron dataset.

list-lassy-treebanks

Preparation for search through the Lassy corpus. Creates a list of the paths for all XML files in LassyLarge.

lassy-search

Searches through Lassy's XML files and extracts instances of relevant verbs.

assemble-vocab

Gets the word frequencies from the Lassy corpus and creates a list of t 10.000 most frequent words as the training vocabulary.

filter-observations

Checks if there are verbs with no training data, and the size of training data for each verb. Also checks if there are training files with no corresponding verb.

create-cooccur-matrix

Uses assembled sentence to create raw cooccurrence counts for each verb-argument combination.

sort-cooccur-matrix

Sorts the cooccurrence data, which vonverts the cooccurrence file to a sparse matrix format with named columns and rows.

weight-cooccur-matrix

Applies PPMI weighting to the cooccurrence matrix.

number-cooccur-matrix

Changes the column and row names into index numbers, which makes processing easier.

reduce-cooccur-matrix

Apply SVD to reduce the number of columns in the cooccurrence matrix from 10.000 to 200.

assess-regression-sample-size

Counts the number of type arguments for each verb, which constitutes training size for regression.

numeraltostring

Defines a class which transforms some numeral expressions into strings with their full Dutch form.

transformargs

Helps to match argument strings with the Tulkens vectors by offering some transformations on argument strings. Includes handling for compound words and diminutive forms.

train-verbs-index

Creates an index for verb training. Checks if argument strings have corresponding vector in the Tulkens dataset. Otherwise, arguments are not included.
For each verb, this lists verb's arguments as the string to retrieve their N vector and the row number corresponding to their holistic vector.

regression-weights

Uses the frequency of each verb-argument combination to calculate their weight in regression training.

train-verbs-testing

Performs Ridge regression to train verb matrices, but does not export the result. Tests out the effect of different parameters and creates a few figures for evaluation.

argv-correlation

Measures the correlation between argument vectors of each verb.

train-verbs-final

Trains verb matrices and exports the result for verbs with a sample size > 400.

compare-phrases

Applies the trained verb matrices to the relpron dataset. Uses a few models of composition and compares results.