Conversion Pipeline RIDGES 4.1

Data formats

The annotations for RIDGES are currently done using Excel files. They need to be converted to the ANNIS import format. As intermediate format, PAULA (http://www.sfb632.uni-potsdam.de/paula.html) is used.

Installation of required software

Currently this conversion only works if you have a Windows operating system and Excel installed. You will need a special Excel Add-In to export the data to PAULA. This Add-In is a modification of the Exmaralda Exccel Add-In from Amir Zeldes (http://www.exmaralda.org/en_exceladdin.html). Follow the instructions at http://www.exmaralda.org/converters/README_exmaralda_io_0.9.8.1.pdf to install the Add-In, but use the file http://korpling.german.hu-berlin.de/ridges/download/v4.1/ridges_io_0.9.8.1.xla instead of the original one. The name of the Add-In will not be “Exmaralda_Io_0.9.8.1” but “Ridges_Io_0.9.8.1”. Additionally the menu entry for the Add-In is not “Exmaralda” but “Ridges_IO”.

Next download SaltNPepper by following the instructions at https://github.com/korpling/pepper#download-and-install. For the conversion of RIDGES 4.1 at least version “SaltNPepper_2014.09.24-SNAPSHOT” is needed.

Download or checkout from Subversion

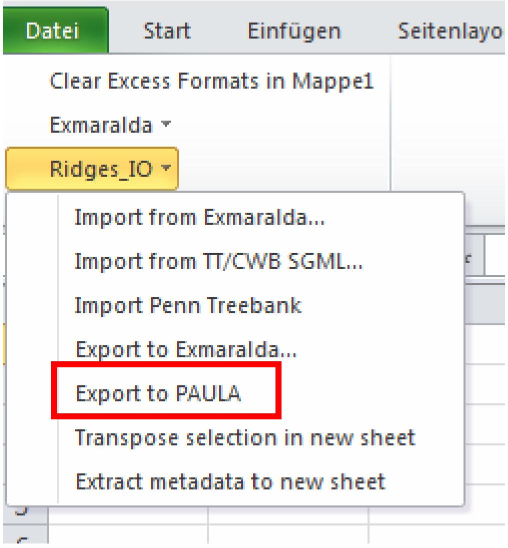
Either download the corpus files (http://korpling.german.hu-berlin.de/ridges/download/v4.1/Excel_with_conversiontemplate.zip) or check them out from our Subversion repository. In the end you should have a data folder with the following sub-folders and files:

- “Excel” → folder containing the Excel files
- “paula” → folder where the paula files will be located
- “relANNIS” → folder where the ANNIS import files will be located
- “relANNIS_template” → folder with some ANNIS import files that are not created during conversion but have to be copied manually
- “paula2relANNIS.pepperparams” and “addorder.prop” → description files for the SaltNPepper converter

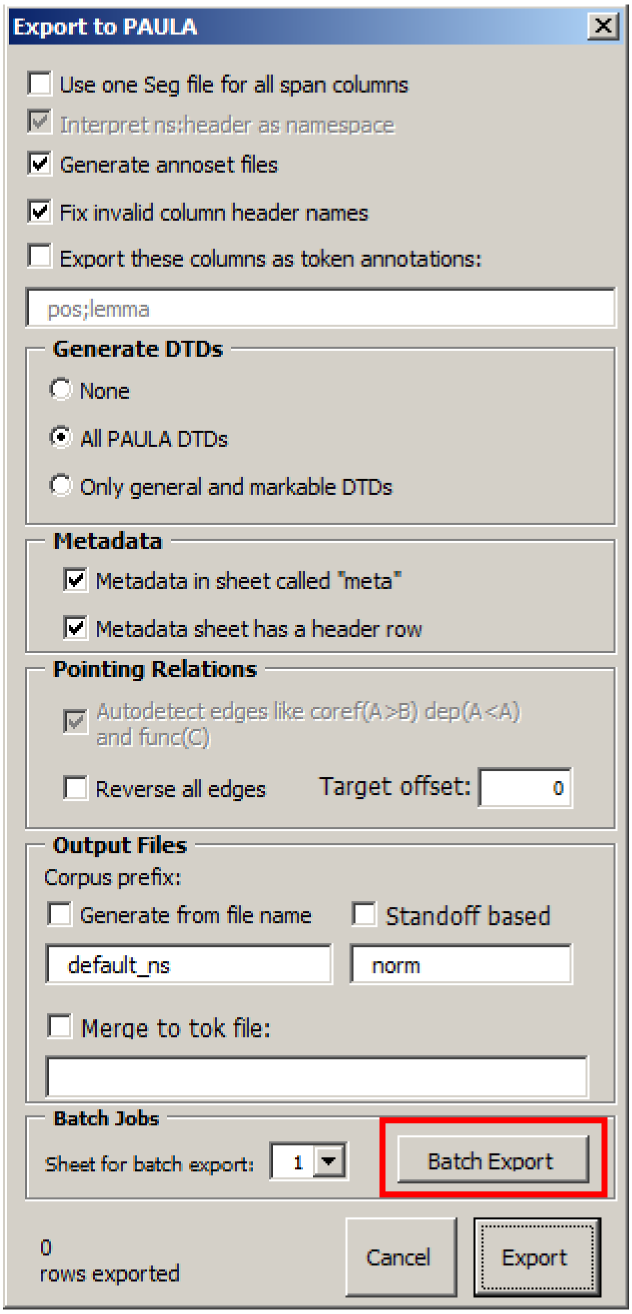
This folder gets the placeholder `DATA_DIR`.

Conversion

1. Open Excel
2. Open the “Add-In” ribbon
3. Choose “Ridges_IO” and select “Export to PAULA”



4. Make sure the settings are the same as in the picture below and click on “Batch Export” to start conversion.



Export to PAULA

☐ Use one Seg file for all span columns

☑ Interpret ns:header as namespace

☑ Generate annoset files

☑ Fix invalid column header names

☐ Export these columns as token annotations:

pos;lemma

Generate DTDs

○ None

● All PAULA DTDs

○ Only general and markable DTDs

Metadata

☑ Metadata in sheet called "meta"

☑ Metadata sheet has a header row

Pointing Relations

☑ Autodetect edges like coref(A>B) dep(A<A) and func(C)

☐ Reverse all edges   Target offset: 0

Output Files

Corpus prefix:

☐ Generate from file name   ☐ Standoff based

default_ns   norm

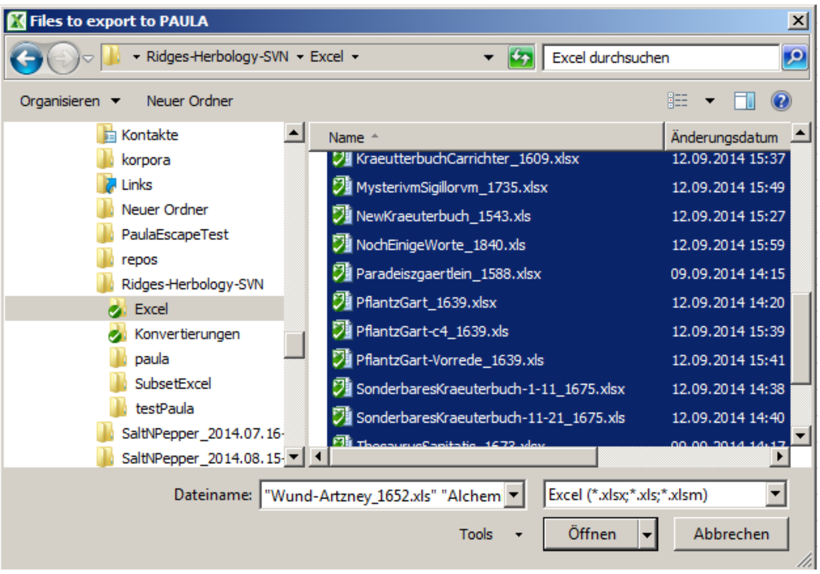
☐ Merge to tok file:

Batch Jobs

Sheet for batch export: 1

0
rows exported

5. Select all Excel files you want to convert in the file chooser dialog and press on “Open”.



Files to export to PAULA

Ridges-Herbology-SVN ▸ Excel

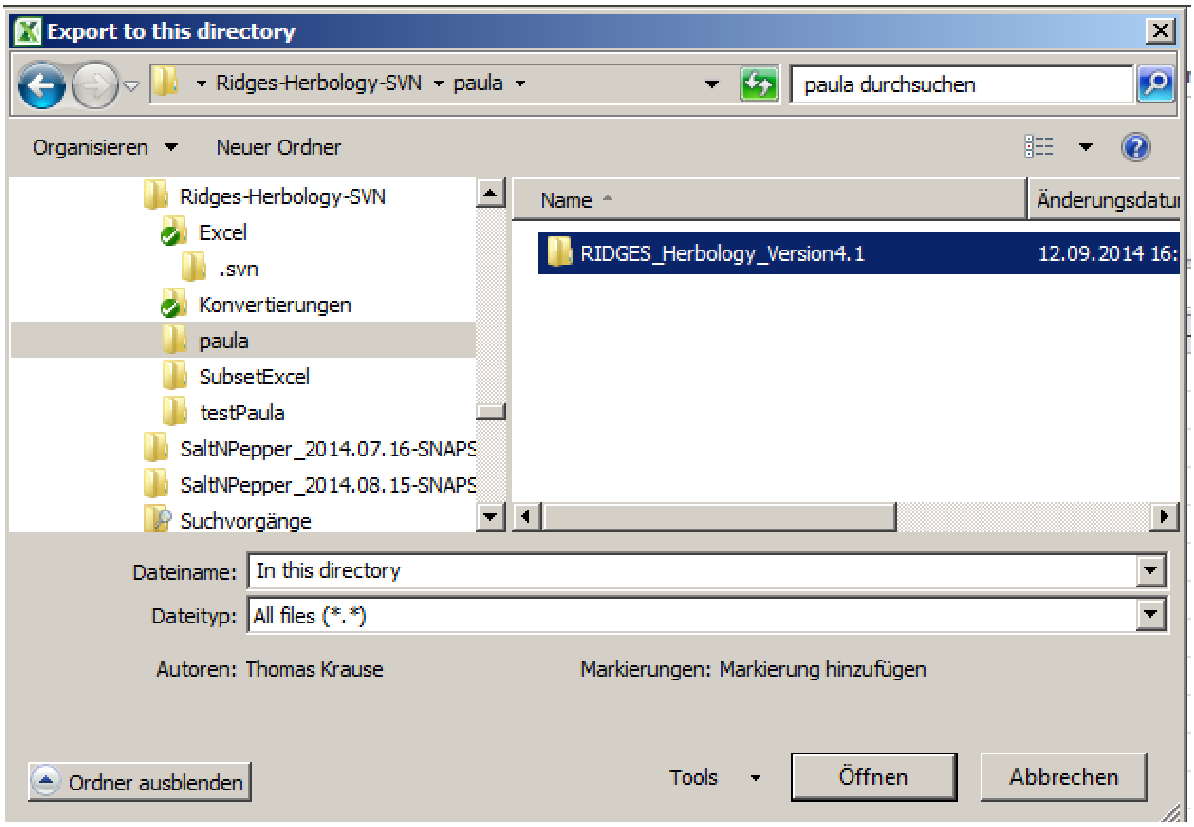
Excel durchsuchen

Name	Änderungsdatum
KraeutterbuchCarrichter_1609.xlsx	12.09.2014 15:37
MysterivmSigillorvm_1735.xlsx	12.09.2014 15:49
NewKraeuterbuch_1543.xls	12.09.2014 15:27
NochEinigeWorte_1840.xls	12.09.2014 15:59
Paradeiszgaertlein_1588.xlsx	09.09.2014 14:15
PflantzGart_1639.xlsx	12.09.2014 14:20
PflantzGart-c4_1639.xls	12.09.2014 15:39
PflantzGart-Vorrede_1639.xls	12.09.2014 15:41
SonderbaresKraeuterbuch-1-11_1675.xlsx	12.09.2014 14:38
SonderbaresKraeuterbuch-11-21_1675.xls	12.09.2014 14:40

Dateiname: "Wund-Artzney_1652.xls" "Alchem

Excel (*.xlsx;*.xls;*.xlsm)

6. Choose the output folder, “paula/RIDGES_Herbology_Version4.1”. The output folder must have been created before.



RIDGES_Herbology_Version4.1 12.09.2014 16:

Dateiname: In this directory

Dateityp: All files (*.*)

Autoren: Thomas Krause

Markierungen: Markierung hinzufügen

7. The RIDGES corpus has some large documents. In order to be processed by SaltNPepper you have to adjust the SaltNPepper configuration. If `PEPPER_HOME` is the installation folder, edit the file `PEPPER_HOME\pepperStart.bat` and change the line

```
java -Xmx1024m -XX:-UseGCOverheadLimit -cp lib/*;plugins/*;
-Dlogback.configurationFile=./conf/logback.xml
de.hu_berlin.german.korpling.saltnpepper.pepper.cli.PepperSta
rter %1 %2
```

to

```
java -Xmx4000m -XX:-UseGCOverheadLimit -cp lib/*;plugins/*;
-Dlogback.configurationFile=./conf/logback.xml
de.hu_berlin.german.korpling.saltnpepper.pepper.cli.PepperSta
rter %1 %2
```

This will set the maximal amount of internal memory to about 4GB. You will need a computer that has at least 8GB RAM for this setting. If you have a computer with more memory you can set the value even higher. Also adjust the file `PEPPER_HOME\conf\pepper.properties` and change the property “pepper.maxAmountOfProcessedSDocument”:

```
pepper.maxAmountOfProcessedSDocuments=3
```

If using the default value 10 too many documents will be hold in memory at the same time which can cause a main memory shortage.

8. Open a command line and execute the following (replace `PEPPER_HOME` with the actual installation directory and `DATA_DIR` with the directory where the files have been unzipped to):

   ```
   cd PEPPER_HOME
   pepperStarter.bat -p DATA_DIR\paula2relANNIS.pepperparams
   ```

9. After the conversion finished, copy the “ExtData” folder, the “example_queries.tab” and the “resolver_vis_map.tab” file from `DATA_DIR\relANNIS_template\` to `DATA_DIR\relANNIS\`.

10. Additionally add the content from the file `DATA_DIR\relANNIS_template\toplevel_corpus_annotation.tab` to the beginning of the file `DATA_DIR\relANNIS\corpus_annotation.tab`. This adds the meta-data for the corpus itself.