Morphilo Documentation

Hagen Peukert*

Abstract

The morphilo toolset consists of three components: Morextractor, Morphilizer, and Morquery. Morextractor commands a reductionistic logic matching a set of affix strings to the given word input by using a simple rule set of the English Morphology. Since this algorithm is highly overgeneralizing, the second programm aims at correcting the overgeneralizations and storing the correct entries in a database. Last, Morquery is a tool to conveniently query the database for all common features encountered in English derivational morphology.

1 Preliminaries: requirements, downloads, and installation procedure

Provided that you have downloaded the *morphilo* package from the LiMA webpage[1] containing – besides this documentation – a folder with three .jar-files (morextractor, morphilizer, morquery), a file named *dbentry*, a *lib* folder with several class files and a SQL-file *InitialEntrySetup*. You should save these files on your hard drive at a location convenient to you.

The algorithm of the software requires tagged corpora in a textfile format although showing the extension *.pos*. If you have only plain texts without any annotations available, you may choose a rule based (montytagger) or probablistic (stanford parser) tagging method to automatically annotate the corpus. In the latter case, you will have to correct the "mistaggings" manually while running the Morphilizer component. Make sure you change the file extension to *.pos*. Usually these taggers will save their results as *.txt*.

To set up the toolset, three steps have to be taken.

1. download MySQL-DBMS and install it

2. set up data base schema and tables

3. update the dbentry-file

[*] The development of the software was carried out at the LiMA excellence cluster initiative *multlingual spaces in urban areas.*

[1] http://www.lima.uni-hamburg.de/subdomain/ect

1.1 Downloading MySQL and Installing the MySQL DBMS

If you have already an MySQL client installed and you have access to a server with sufficient disc space, you can skip this step. If this is not the case, you should download the MySQL-GE and the MySQL Workbench GUI-tool from the MySQL webpage[2]. If you choose to run the database on your hard drive, pick the community server download, klick the downloaded file, and follow the instructions given there.

Open the Workbench and set up a “new server instance” either to your hard drive or to a remote server. For the latter, you have to specify the IP-address and login details. Note down the IP-address or *localhost* respectively, the username as well as its password. Running it on your hard drive, the presets should suffice, but make sure that the server is switched on when using it (This is usually done in the system's setup). Then klick on “new connection” and connect to the just created instance.

1.2 Setting up data base schema and tables

Open the Workbench, choose “Open SQL Script...” from the file menu, and browse to the morphilo package. Select “InitialEntrySetup.sql” and confirm your choice. If you do not like the preset database name *morphilodb*, you have to update the SQL script. The easiest way to do this is to use a text editor and to exchange each sequence of *morphilodb* by your choice of a database name. Remember the name of the database. Finally, choose from the *Query* menu, *Execute (All or Selection).*

The script should run through without error. If it does not, make sure you have not only selected parts of the SQL script. In case you have made changes to the database name, check if you changed all of the old name tags correctly.

1.3 Updating the entryDB-file

The last step is to configure the “dbentry” file. The “dbentry” file should stay in the same directory as the .jar-files. If you miss this step, you are asked at the start up of the Morphilizer or Morquery tool to do so. The configuration is very simple and quickly done, but you should pay attention to the order of each item you will have to put in.

Open the file with a simple text editor. The first item to occur in this file is the IP-address or name of the server. For hard drive installations put “localhost” or “127.0.0.1”. The second line must have the name of the database. Preset is *morphilodb.* If you have changed it, put down the new name. The third line should contain your password and the last line specifies

[2] http://www.mysql.de/downloads/

your username. Nothing else should be in the file. Save and close it. You are now ready to start the morphological analysis.

2 Description and Usage

2.1 General Procedure

After the installation procedure and database setup, you are ready to use the software. The software package is designed to analyze derivational morphology in the course of time using large corpora. It comprises several steps.

First, the corpus data have to be pre-analyzed, that is, conveyed to a form that is easier to process in terms of derivational morphology. This is done by Morextractor. It assigns additional tags to the words of a corpus revealing possible affixes and compounds. Depending on the speed of your system, this may take up to 30 minutes for very large corpora, since the enumerated lists of suffixes and prefixes contain all allomorphemic variants, which add up to more than some $1,000$ items.

The second step involves manual work. The Morphilizer software will support you in correcting the automatic analysis wherever words are oversegmented. Each word has to be analyzed only once for each time period. All other occurrences are automatically readjusted and saved in the database.

Once you have finished your analysis you can query your data. For all queries start the Morquery tool by double clicking the respective jar-file.

2.2 Morextractor

Double click the morextractor.jar file to start working on diachronic morphological analyses. A simple platform (fig. 1) will enable you to browse to the folder of your tagged corpus files. Notice again that only *.pos*-files will be selectable. All files of the folder are processed. If you wish to process single files, you have to copy them to separate folders and select them one after another. This ensures that you will not double-pick one file and thus prevent skewing the quantitative analyses. Also select a destination where a file system consisting of several folders will be set up. The folder in which the preprocessed files are written to is called “wordlist”. You may also choose to restrict the analysis to one specific word form by checking the respective box. Otherwise choose the *all* check box. The text field at the bottom of the tool prompts you with possible errors, data inconsistencies, or the determination of the automatic analysis. Once all calculations are finished, you can proceed with finetuning these results using the Morphilizer tool.

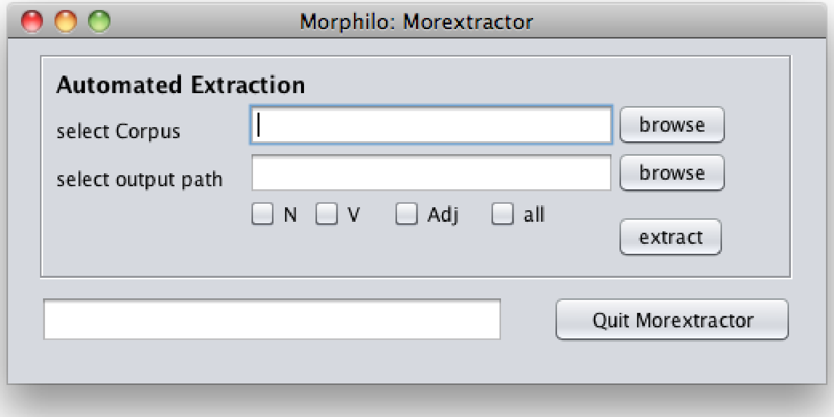


Figure 1: Morextractor Screenshot

2.3 Morphilizer

Figure 2 displays the user interface for the Morphilizer component. The first step is to choose the preprocessed file as described in the previous chapter. You find the file in the *wordlist* folder. Select it in the *file selection* section and choose a corpus in the *corpus specification* section from the drop down menu or by individual specification in the respective text fields. Then press *start*. Morphilizer will now check all entries for existence in the database. If so, the respective database tables are updated and the entries are deleted from the *wordlist*-file. Depending on the size of the *wordlist*-file this may take several minutes.

Once this mapping is done, the first entry is displayed in the *Derivation* section on the left side of the user interface. You have to make a choice on whether Morextractor assigned the correct affix or composition information. Use the radio buttons to correct false mappings or press the *skip*-button to exclude the entry in the analysis. All skipped items are saved to an extra list, which can be analyzed at a later point in time.

It is recommendable to consult the online OED dictionary to make the correct choices since the line between composition and affixation is far from clear-cut. This is especially true for diachronic data (e.g. the Middle English Period). From our experience, it is best to place the OED online interface right next to the Morphilizer tool. Whenever uncertainties in the analysis arise, you can check the time-usage pattern in the ethymology section of the OED. When your clear on the right affixation, press *OK*. The entry is now stored in the database. For the given time, you will not have to do the analysis of this item again. Once the database grows larger in size, there will be only little entries left that need an extensive research on its affixation.

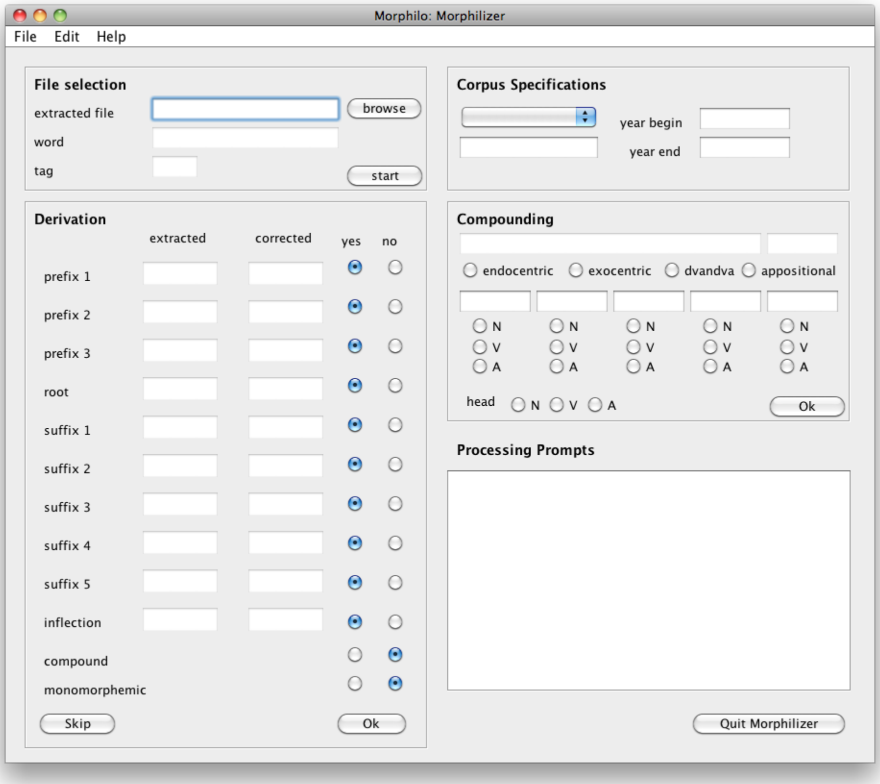


Figure 2: Morphilizer Screenshot

PENN Parsed Corpora traditionally encode composita. If this is the case, they are automatically put in the *Compounding* section. Unfortunately the underlying definition of what compounds are is not convincingly put into practice in terms of the corpus annotations. Many composita are missed. Select the *compound* radio button if a word is really a compound but not annotated as such by the corpus annotations. The compositum will then be displayed at the right hand for further analysis. You may have to manually write each free morpheme of the compound to one of the specified text fields, select its word form, specify the head information, and make a choice on the compound type (endocentric, exocentric, dvandva, or appositional). After pressing the *OK* button here, each of the specified morphemes can be analyzed for its derivational structure in the *Derivation* section.

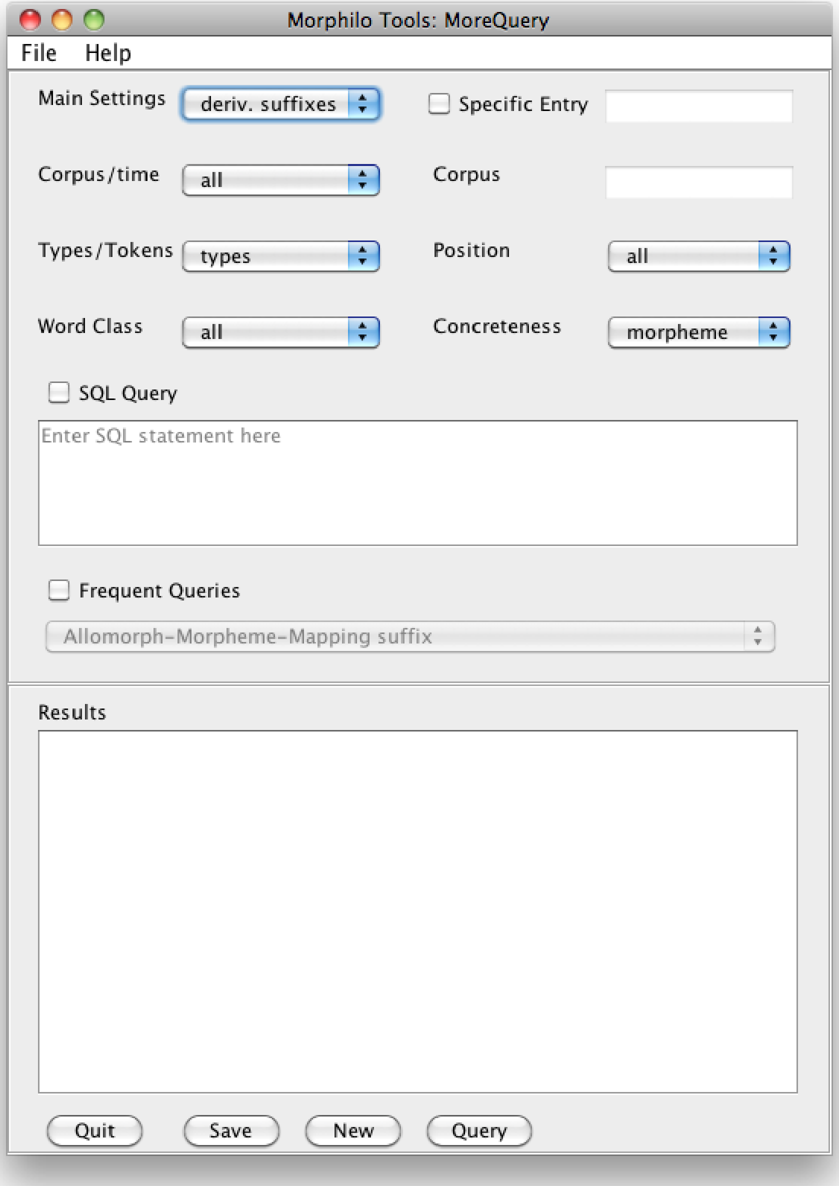
Last, the *Processing Prompts* section gives you a feedback on the state and success of all major operations carried out (e.g. whether a word was successfully written to the database or failed to do so).

2.4 Morquery

The Morquery component (fig. 3) is a query tool. It eases to enter long SQL statements by selecting a combination of drop down menus and check boxes. You just have to choose the information of interest from the drop down menus and the information is translated to an SQL query and executed. The menus provide you with types and tokens of all affixes, compounds, words or subsets thereof, either morphemes or allomorphs as well as its position and head information. It is also possible to enter full SQL statements in the SQL query section. The most common queries can also be selected in a drop down menu. The *Results* text area displays all querying results. They can be saved as a textfile by hitting the save button.

3 Continuous Improvement and Testing

The toolset at hand is a first implementation of an use case. The final design in terms of *look and feel* of the graphical user interface as well as the analyses and design of data structures and classes will be implemented after a prolonged testing phase. We like to see if the toolset is useful in the present setup to stand the requirements of diachronic morphological analysis. Therefore, we like to encourage all user to give us an honest, but fair feedback on the software so that we can use this knowledge to improve its handling, design, and usage requirements.



Morphilo Tools: MoreQuery

Main Settings deriv. suffixes

□ Specific Entry

Corpus/time all

Corpus

Types/Tokens types

Position all

Word Class all

Concreteness morpheme

□ SQL Query

Enter SQL statement here

□ Frequent Queries

Allomorph-Morpheme-Mapping suffix

Results

Figure 3: Morquery Screenshot