Heartsome Europe GmbH
Fine Translation Tools

XLIFF TRANSLATION EDITOR | TMX EDITOR | EXTR

User Manual Araya Bilingual Term Extraction

Description how to use the Araya Terminology Extraction Tool

Bilingual Extractor

- The bilingual extractor is a simple to use, but efficient tool to generate automatically term pairs from translated documents (TMX files)
  - A term pair consists of a source and target term
  - A term can consists of several words
- These term pairs can be used to create a new terminology or add new terminology to an existing terminology database.

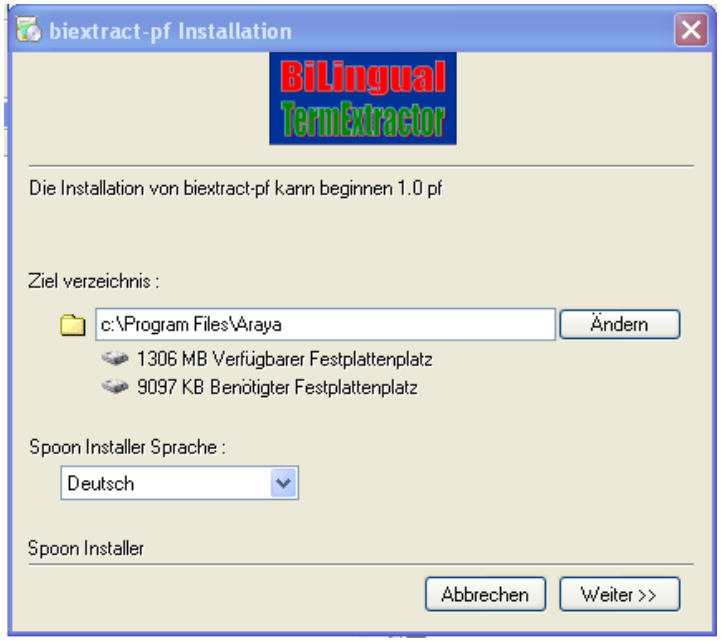
Versions

- The Extractor was developed by Heartsome Europe GmbH and LNE International.
- It is available as
  - Single user version.
  - It is part of the Araya Server translation tools.

Short introduction how to extract

- Extract Terms
  - File -> Extract Bilingual Terminology from File
    - (Option: Open after extraction)
- Check extracted terms
  - Mark correct translation as „validated“
- Export terms
  - Export validated terms …

Installation



- Installation goes into directory **c:/Program Files/Araya**. It is recommended not to change this as all initialization files map towards this directory.

Starting Araya Extraction tool

- Go to directory:
  c:/Program Files/Araya
  Start: BiEdit.exe
- Or double click :

The Extraction Approach

- Based on a TMX file all possible relevant term pairs are computed. This is based on a statistical approach which determines the frequency of source and target terms.
- TMX = XML Exchange format for translation memory databases

Segment

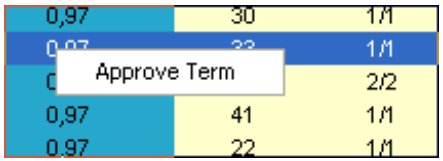
- Extraction is based on segments which are stored in a TMX file.
- A segment can either be a sentence or a whole paragraph.
- Formats in TMX files are ignored.

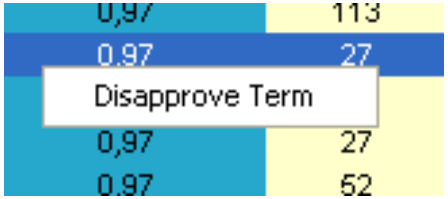
Evaluating and Validating

- Each found term pair is associated with a quality measure.
  - 2. column of the extraction table
  - Value is between1,0 (highest probability that the term pair is a translation) and 0,5 (lowest probability that the term pair is a translation)
- Terms can be validated as correct translations.
  - Last column of table
    - Approved = checked = validated
    - Unapproved = not validated
- Validated terms can be exported.

Validating a Term Extraction Pair

- Select line with term extraction pair
- Validate = approving using
  - Double click term pair
  - Right mouse click
- Remove validation mark using
  - a double or
  - right mouse click



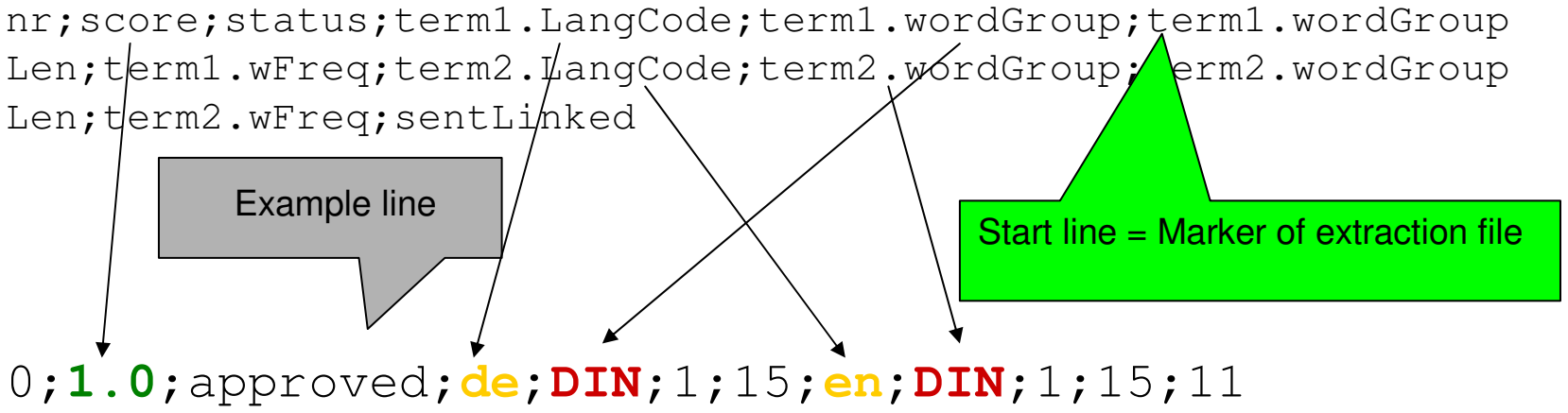


287	0,97	113	1/1	764	787	Kyoto	kyoto	unap...
288	0,97	27	1/1	32	30	Frettchen	ferrets	appr...
289	0,97	27	1/1	27	29	Frischenschlager	frischenschlager	unap...
290	0,97	27	2/2	29	30	El Salvador	el salvador	unap...
291	0,97	52	2/2	130	130	Sierra Leone	sierra leone	unap...

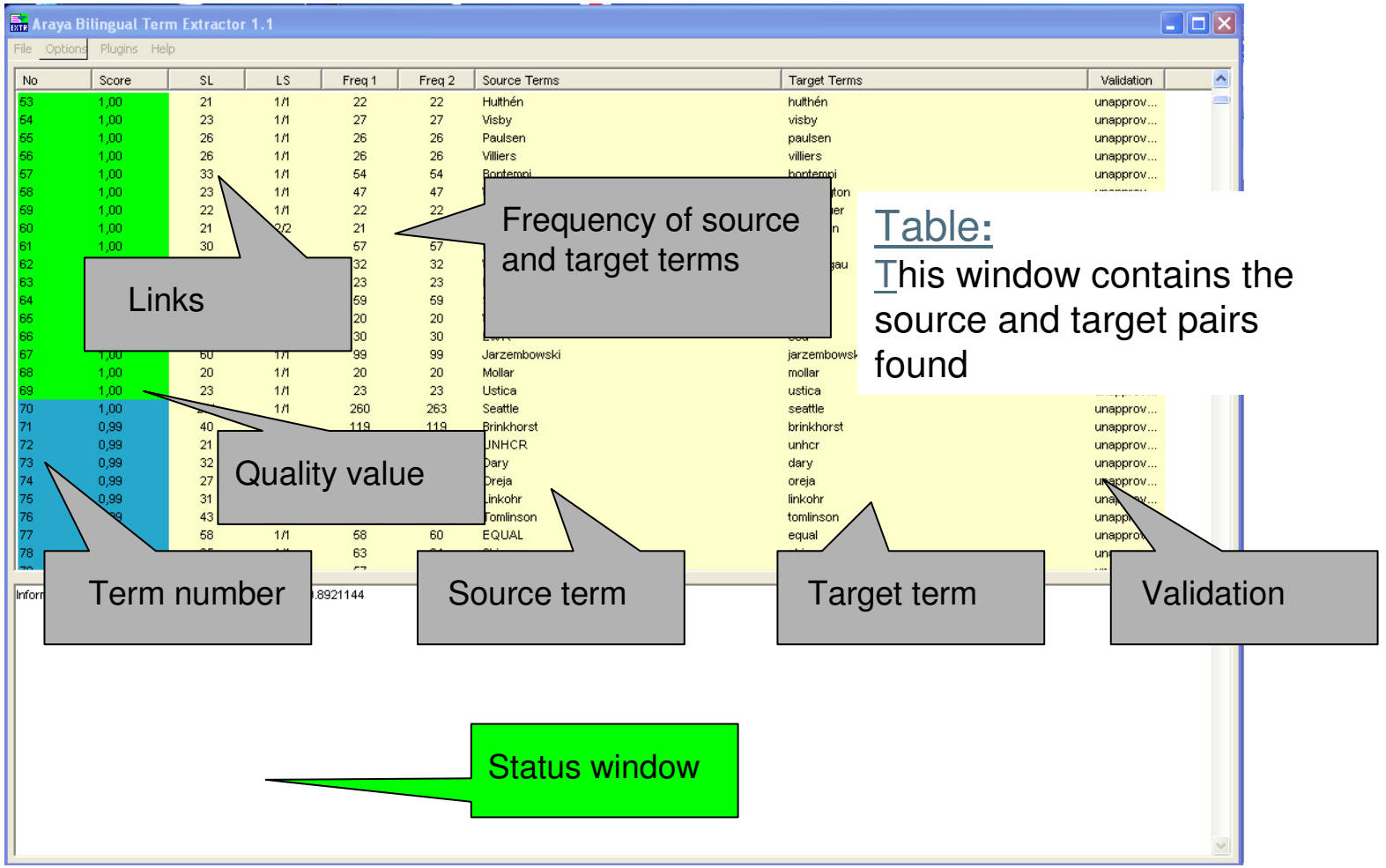
Validated terms appear in green

The Extraction file

- The extraction file has the following format
- Encoding format is UTF-8!



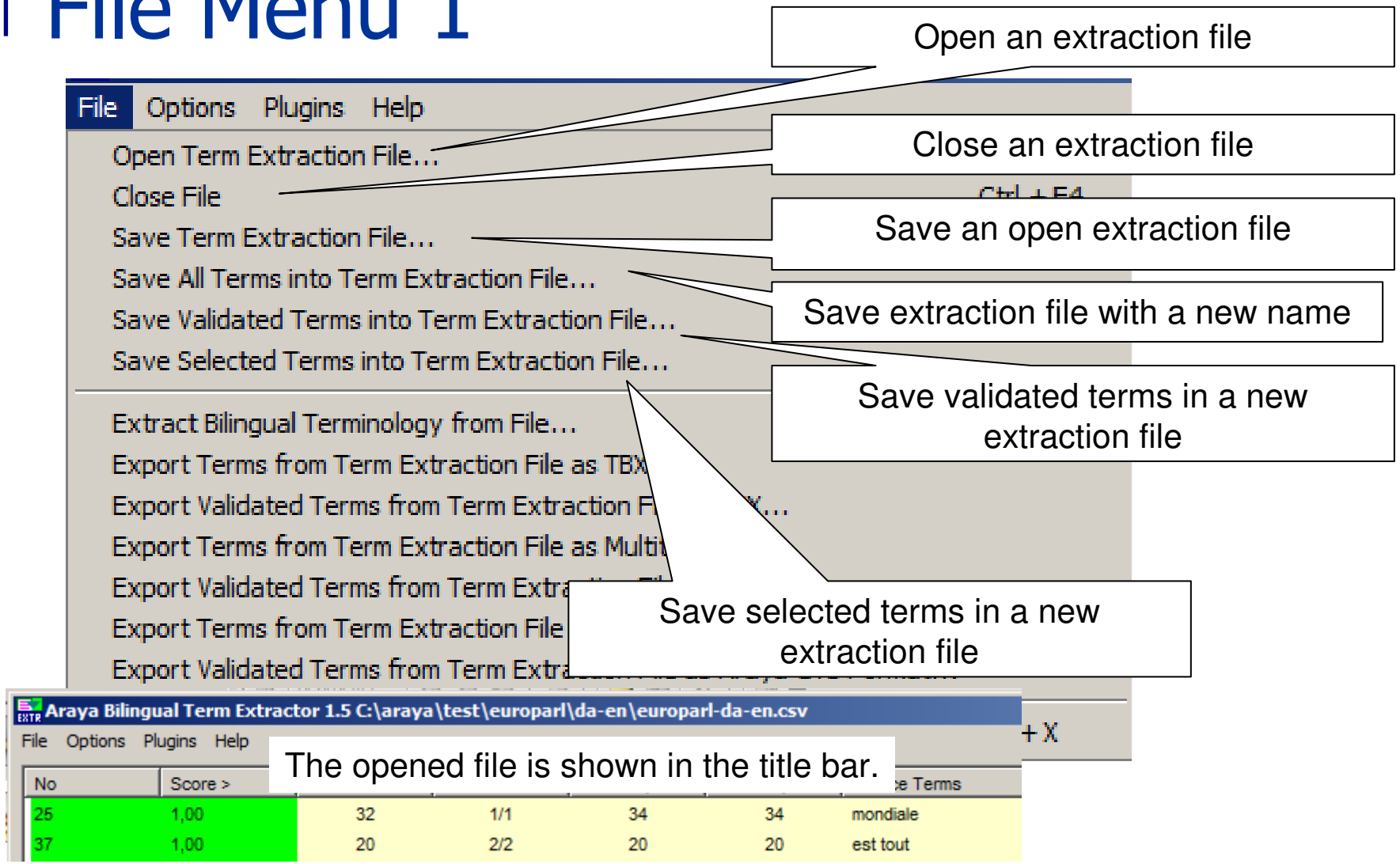
Extraction User Interface



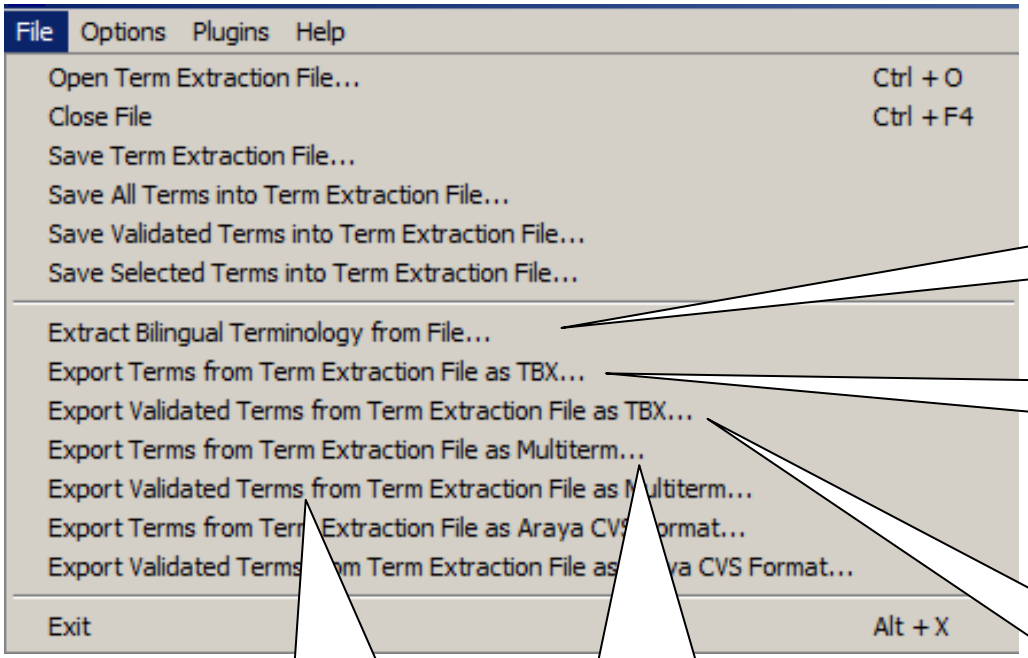
Columns

- Value
  - Statistical measure that the source and target term are translation of each other (quality measure)
- SL
  - Number of segments where both source and target term appear in.
- Freq 1
  - Number of segment where source term appears in
- Freq 2
  - Number of segment where target term appears in
- Source term
  - The source term
- Target term
  - The translation of the source term
- Validation
  - Check box, for marking correct term pairs

File Menu 1



File Menu 2



Extract term pairs from a TMX file

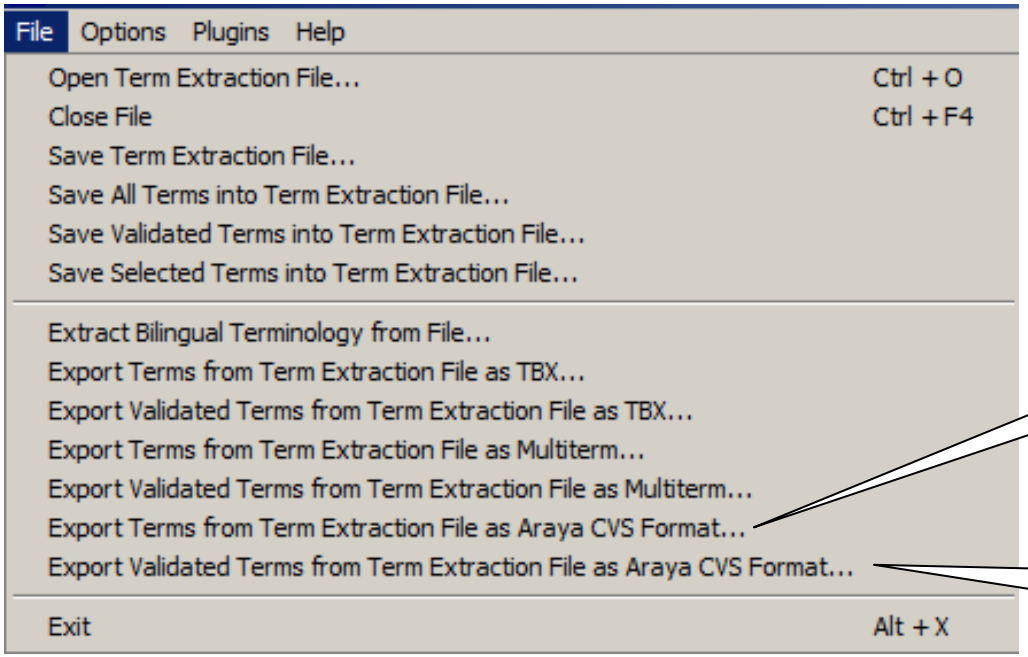
Export terms from open extraction file into TBX Format

Export validated terms from open extraction file into Multiterm Format

Export terms from open extraction file into Multiterm Format

Export validated terms from open extraction file into TBX Format

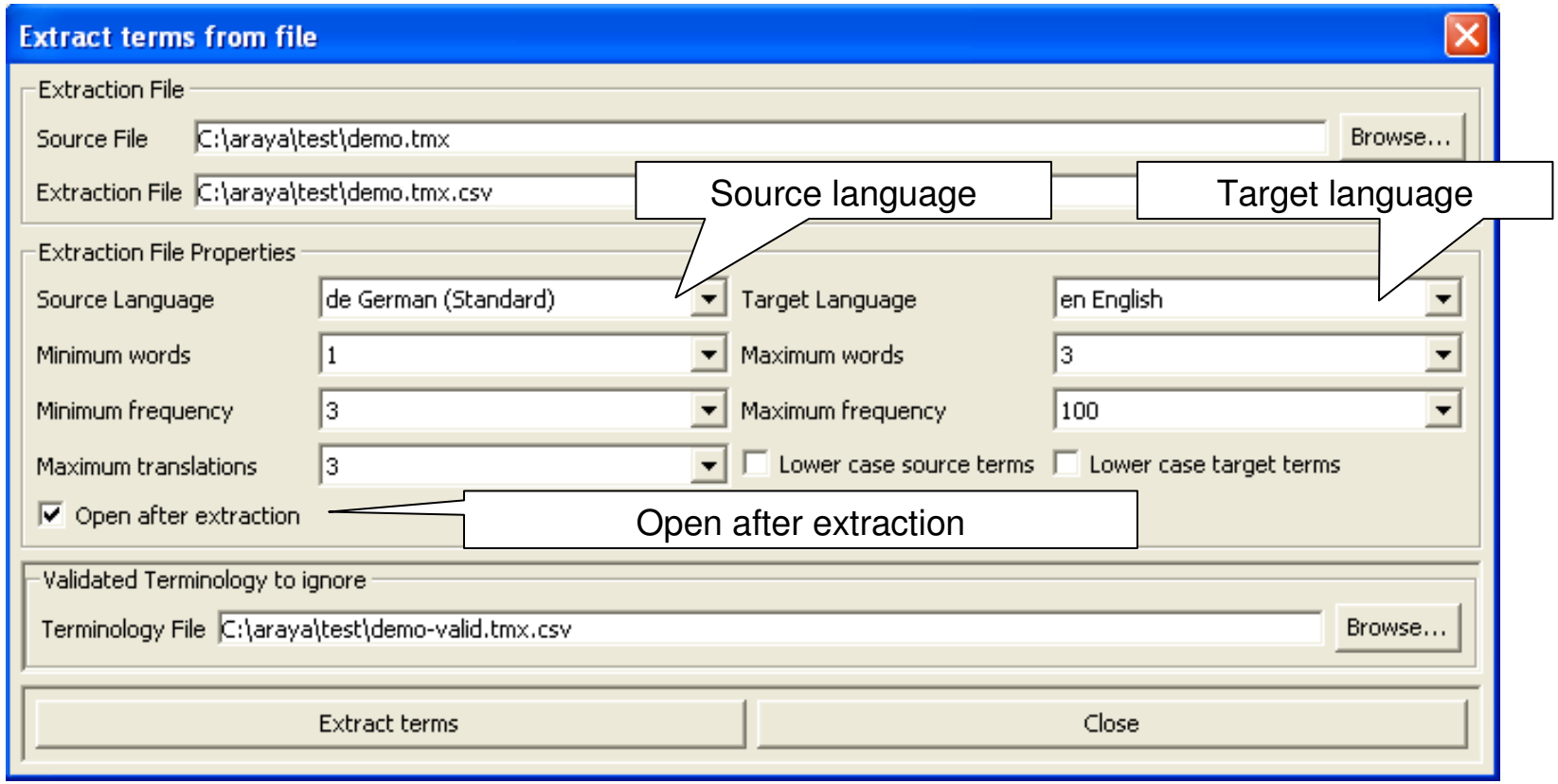
File Menu 3



Export entries from open extraction file in Araya CSV Format

Export validated entries from open extraction file in Araya CSV Format

Extracting Term Pairs from TMX File



Extraction Parameters 1

- Minimum / Maximum words
  - Controls how many words should be contained min/max in the found term pair
- Minimum / Maximum frequency
  - Controls how often the term should appear min/max for the found term pair
- Maximum Translations
  - Controls how many translation should be found at maximum
- Source/Target terms in lower case
  - Controls if source and/or target terms should be converted to lower case

Extraction Parameters 2

- Validated terminology to ignore
  - If a terminology extraction file is specified here, all terms which are marked as “validated” will be ignored.
  - Thru this know translations are ignored.

Exporting

- Exporting can be done in different formats.
  - TBX
    - Name of extraction file + „.tbx“
  - Multiterm (^{TM} of Trados/SDL International)
    - Name of extraction file + „.multiterm“
  - Araya CSV
    - Name of extraction file + „araya.csv“
  - Character encoding is always UTF-8
- Either all or only the validated entries can be exported
- In addition the selection filter (Options -> Export Score Filter) controls the exported terms
  - Depending on chosen value only the term pairs with a minimum score get exported (e.g. score higher than 0.6).

Araya CSV Format

- Araya CSV Format contains the languages in the first line followed by the extracted terms

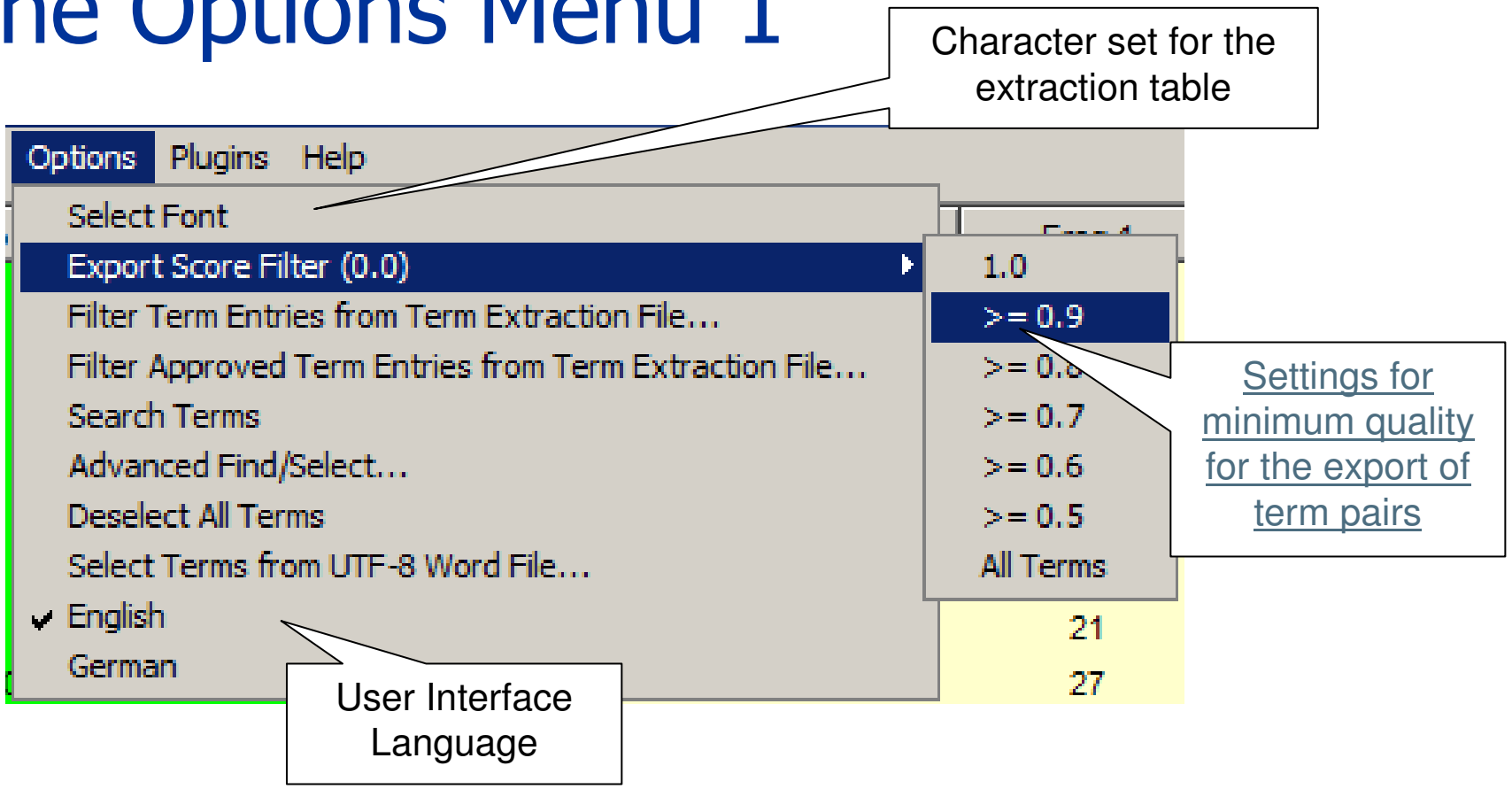
Beispiel

```
de;en
Anschlussplan;Connection diagram
DIN;DIN
Dr;Dr
```

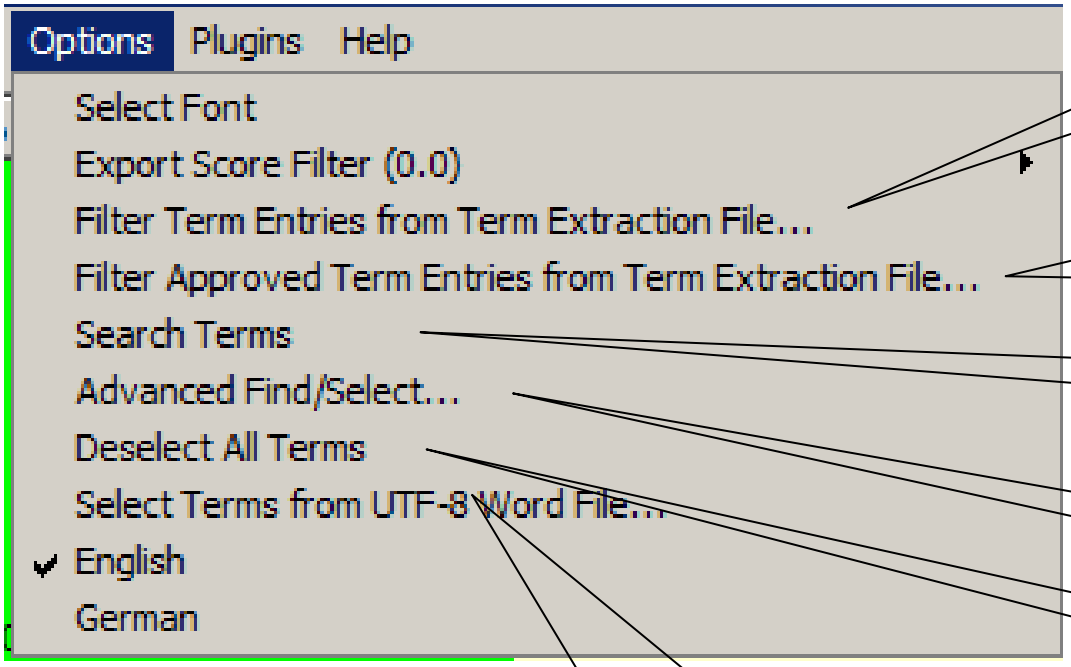
Languages separated by ;

Extracted terms separated by ;

The Options Menu 1



The Options Menu 2



Remove all term entries from the term table contained in term extraction file

Remove all term entries from the term table contained in term extraction file which are approved

Search Terms in the extraction table

Complex search for based on source and target term

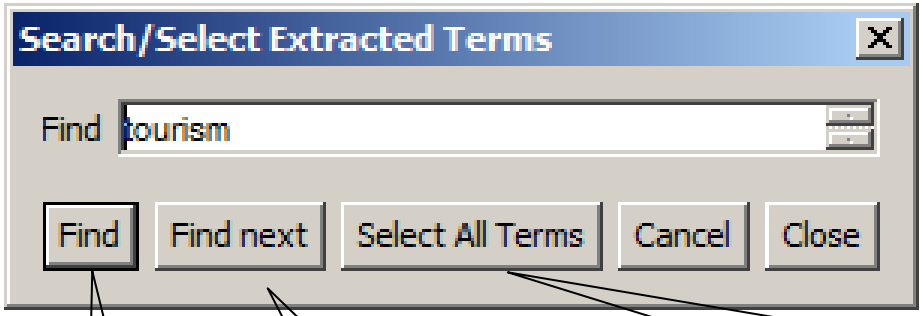
Deselects all selected in the extraction table

Select Terms based on a word list in a file

Filter Term Functions

- The filter functions filters all those term entries which are contained in another term extraction file.
- The identical terms are removed from the term table.
- Depending on the chosen filter method the approved or all entries are used from the specified term extraction file.

Search Term Functions

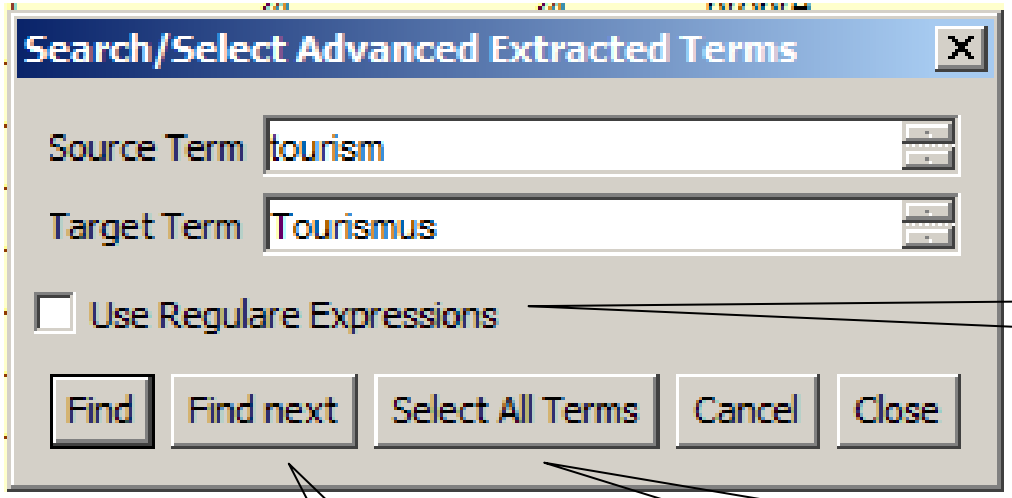


Search this term and select all the matching entries in the table. The selected term candidates can then be saved with “File -> Save Selected Terms into Extraction File…”

Find the next matching term

Start searching this term

Advanced Search Term Functions



Use regular expressions for searching term entries

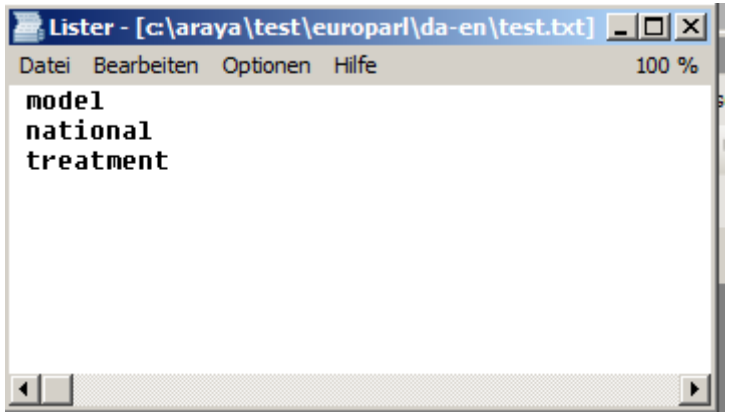
Search this term and select all the matching entries in the table. The selected term candidates can then be saved with “File -> Save Selected Terms into Extraction File…”

Find the next matching terms

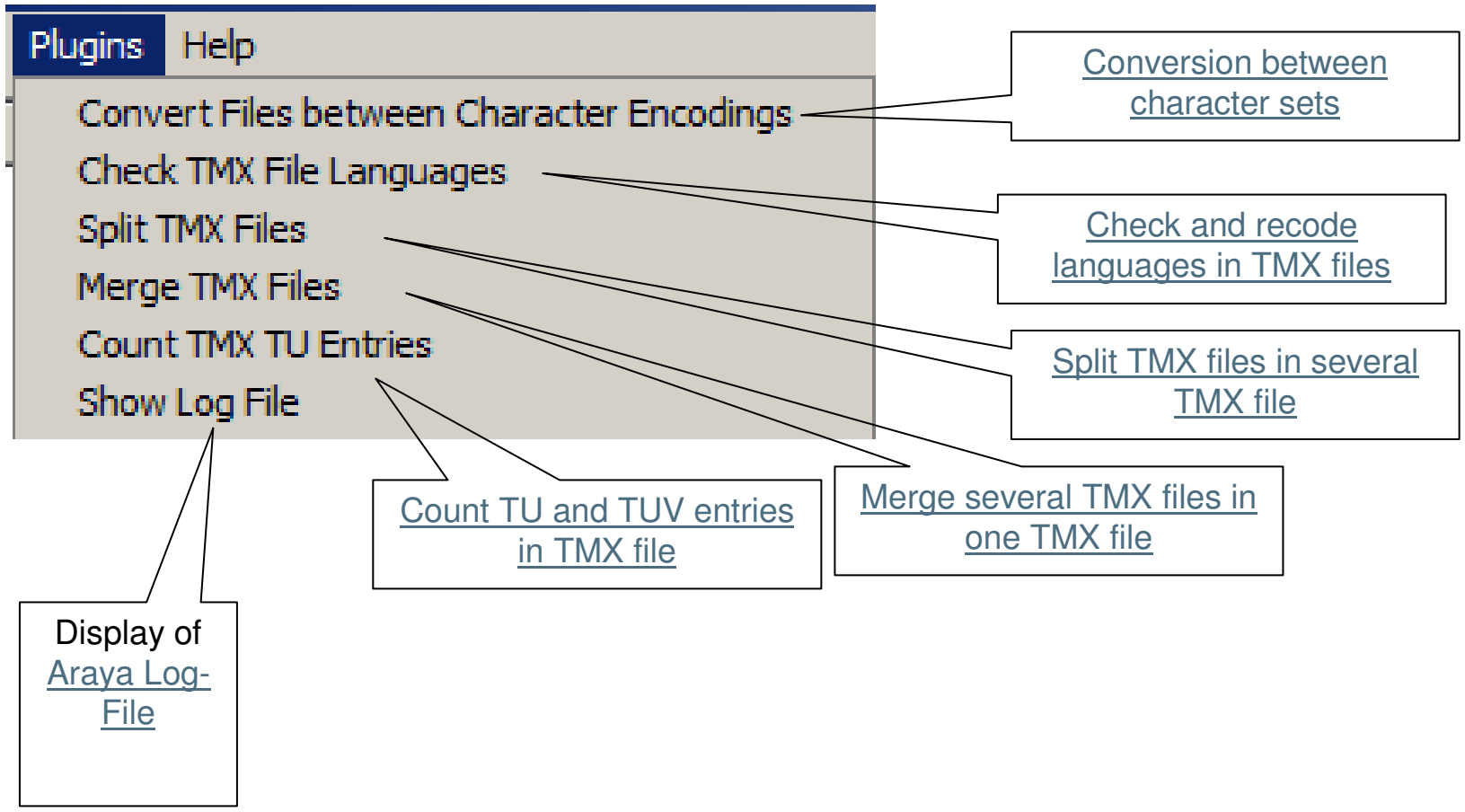
Start searching these combination of source and target terms

Select Terms Based on a Word List

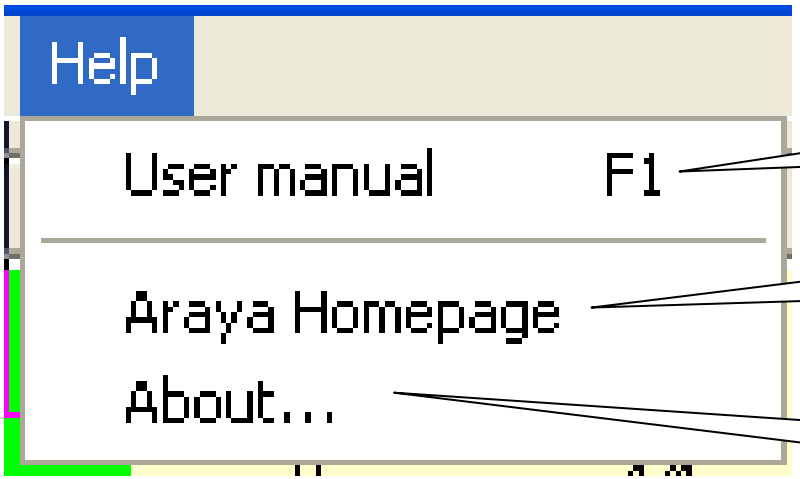
- The filter functions selects all the matching terms from an UTF-8 encoded file. The filter search the given string in both source and target language string
  - It also matches substrings
- The selected term candidates can then be saved with “File -> Save Selected Terms into Extraction File…”
- Each word (string) in the file must be written on a single line.



The Plugins Menu



The Help Menu

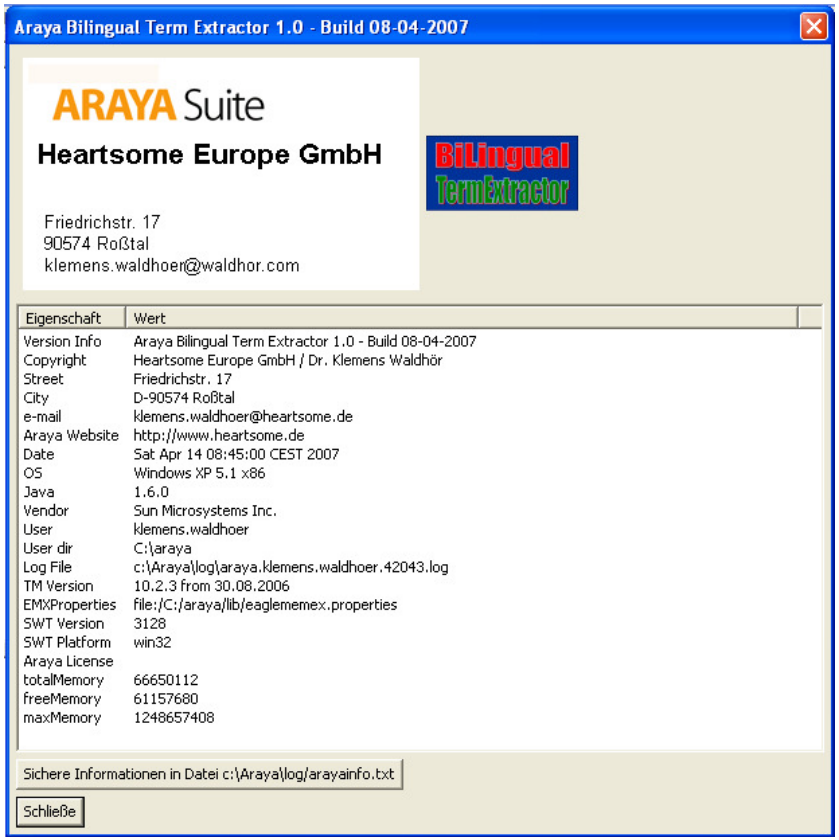


Display user manual

Araya / Heartsome Homepage

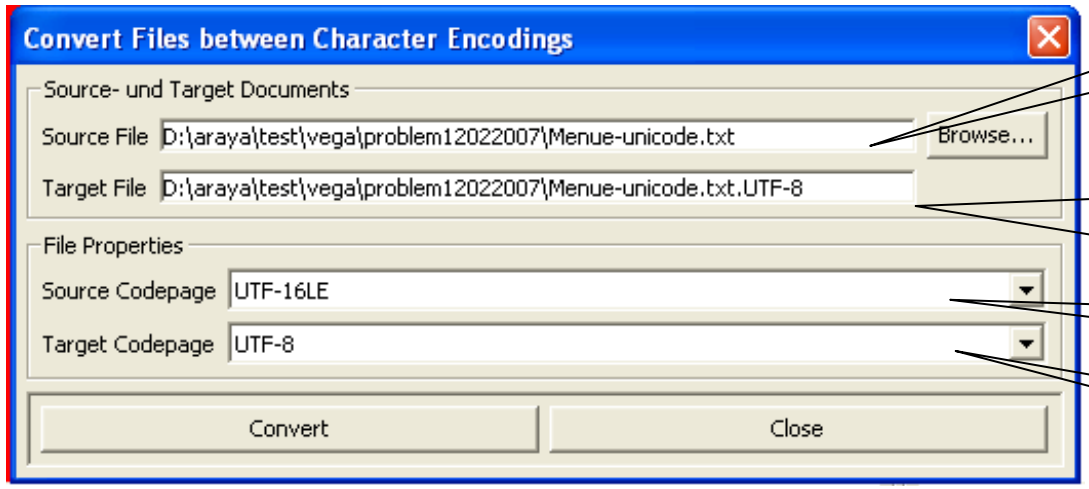
About the bilingual extraction

About the Extraction



- Basic information about the Araya configuration is shown here.
- This can be important for error diagnosis
- Using „Save…“ the information is stored in the specified file.

Converting Files between different Character Encodings



Source file name

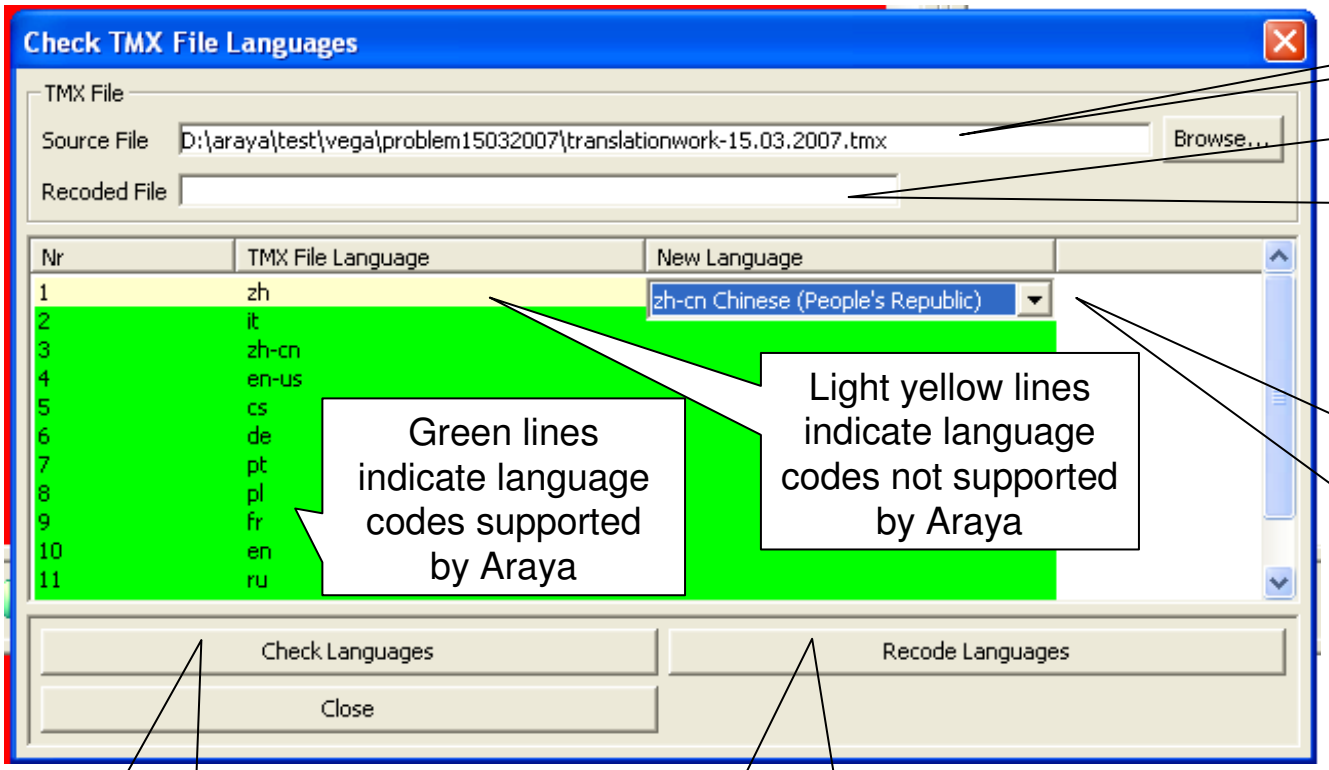
Target file; will be created based on source file name and as extension the target encoding character set

Source encoding character set

Target encoding character set

This function supports converting files between different character sets. Depending on the target character up to two files are written. If the target file is a UTF-8, 16, 32 or UCS file, two files are written. The second file with the extension „.**nobom**“ is a copy of the first target file, the only difference is that the BOM (Byte Order Marks) are removed from this file. This file should be used for importing, esp. when an import of an UTF-8 file is done in Araya, as the Java reading functions for UTF-8 does not over read the BOM characters. This could lead to problems when reading normal strings from those file as the BOMs are read as normal characters resulting in invalid entries.
(This is a known bug in Java UTF-8 file reading, but will not be corrected by SUN!).

Check and Recode TMX Files



TMX file name

Recoded TMX file name; Consists of tmx file name, any ".tmx" extension removed and ".recoded.tmx" added

Language codes supported by Araya Combo box appears only in case language codes in TMX file not supported by Araya

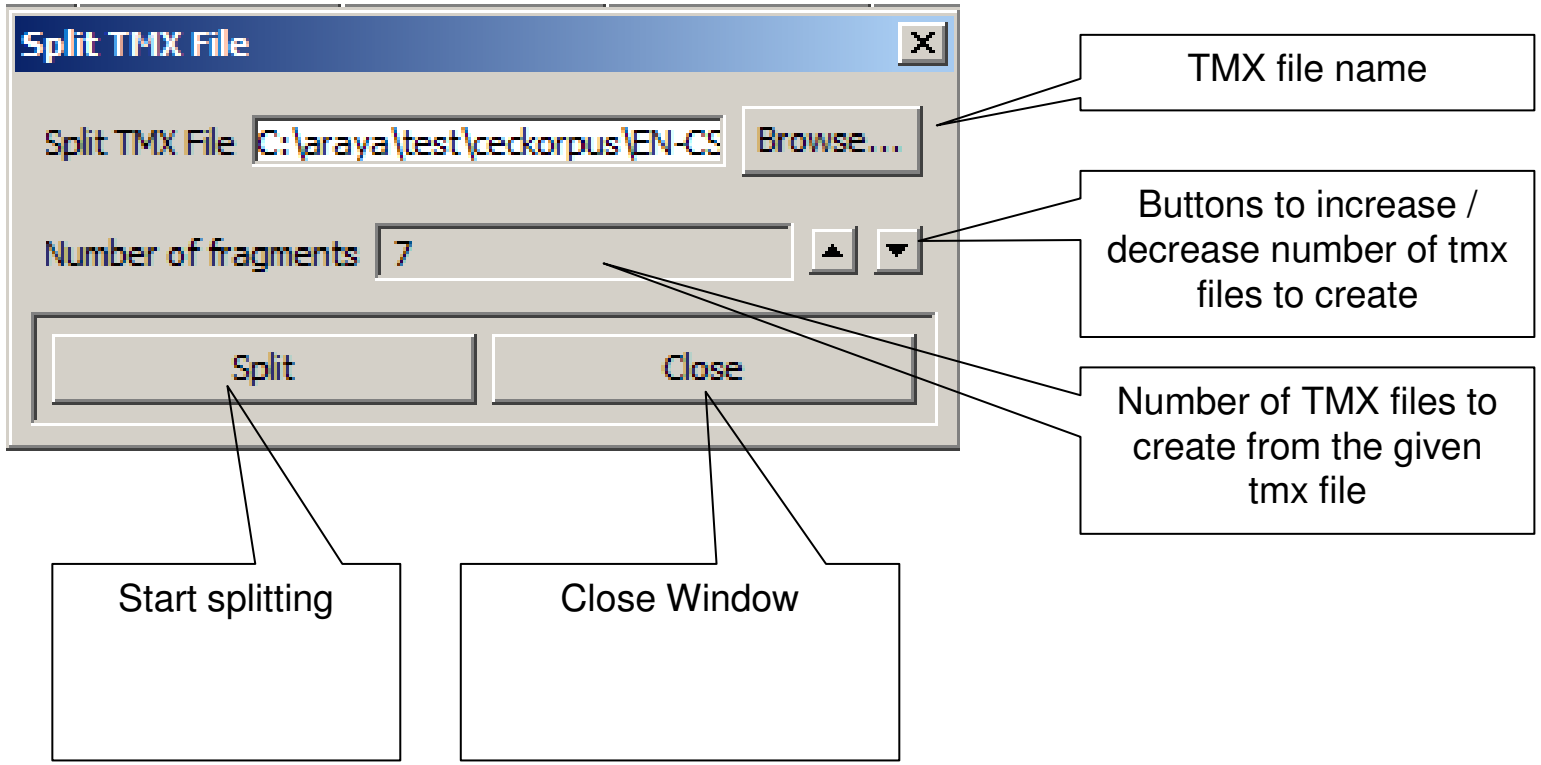
Green lines indicate language codes supported by Araya

Light yellow lines indicate language codes not supported by Araya

Start checking languages contained in TMX File

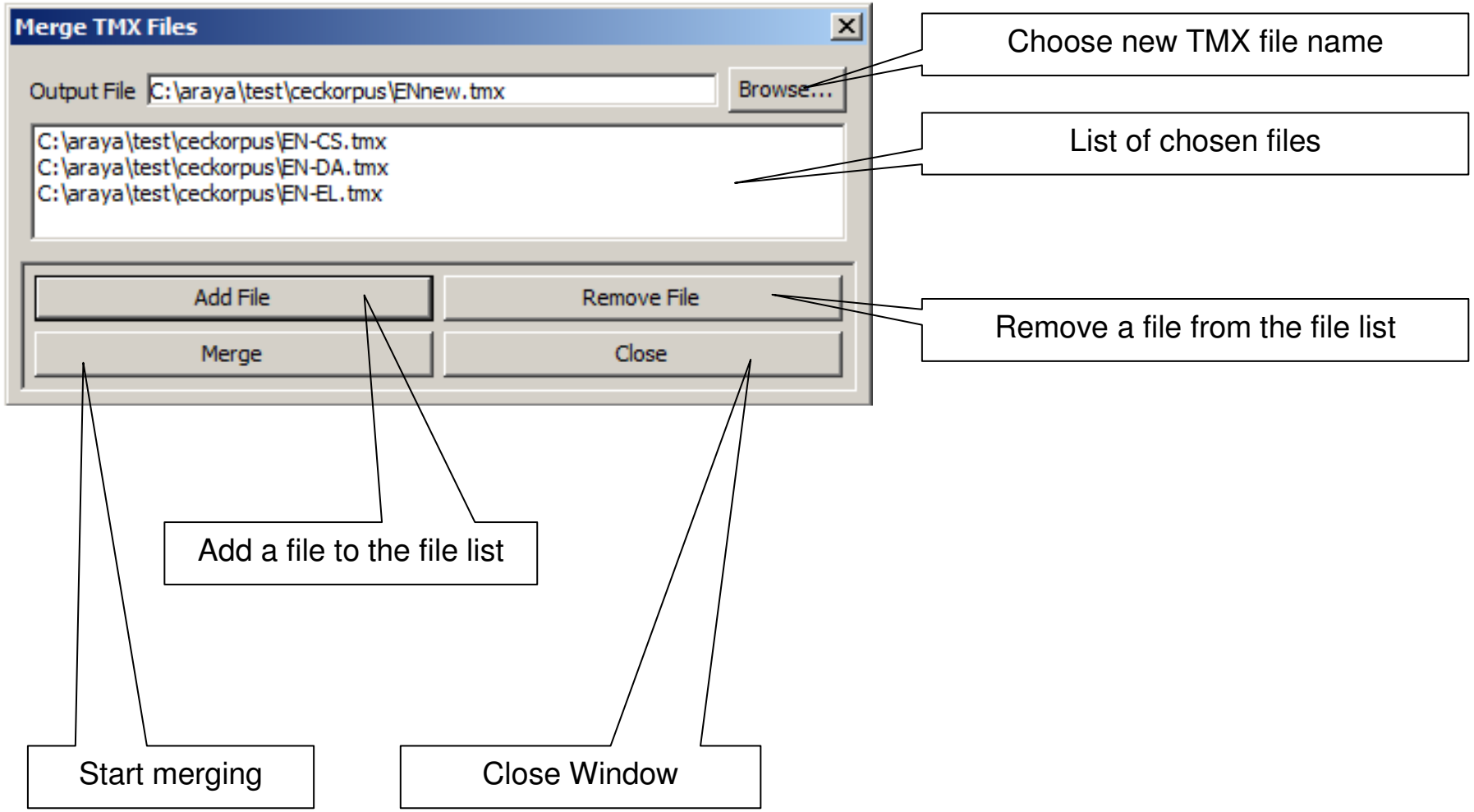
Recode languages using new language for the specified TMX File language

Split TMX File

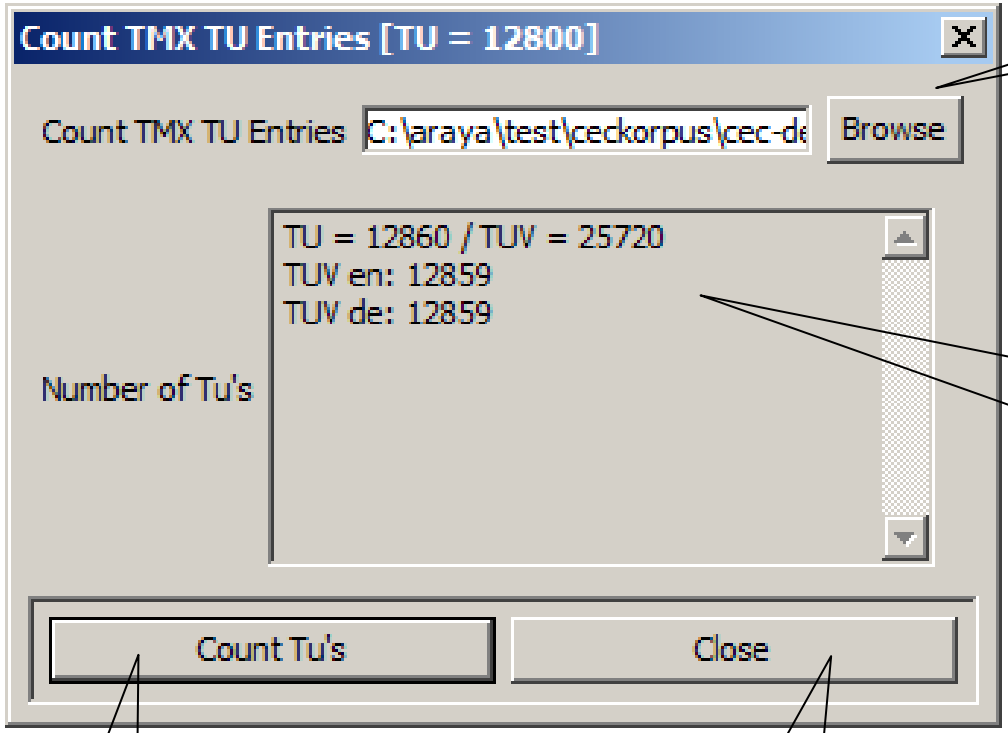


The created tmx files are composed of the original tmx file name plus a number from 1 till the number of chosen tmx file

Merge TMX Files



Count TUs/TUVs in a TMX File



Choose TMX file to analyse

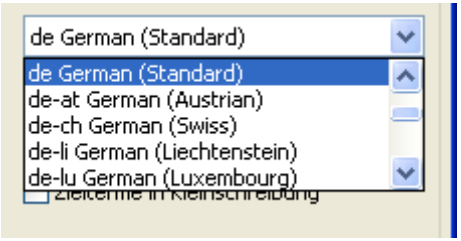
Result:
First line: Number of TUs and TUVs
Following lines:
Number of language specific TUVs

Start counting

Close Window

Adding Language Codes

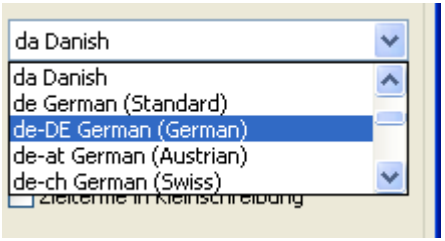
- A set of predefined language codes is defined in the file „ini/lancodes.txt“.
- Additional language codes can be added by modifying this file.
- Here is an example:



Adding de-DE: This requires to add a line like that – where = separates the name displayed an the actual language code

German(DE)=de-DE

Galician=gl
German (Standard)=de
German(DE)=de-DE
German (Austrian)=de-at
German (Liechtenstein)=de-li
German (Luxembourg)=de-lu
German (Standard)=de
German (Swiss)=de-ch



Impressum

- Heartsome Europe GmbH
- Friedrichstr. 17
- D-90574 Roßtal

- Email: info@heartsome.de
- www.heartsome.de
- © 2007, 2009 Heartsome Europe GmbH