Heartsome Europe GmbH
Fine Translation Tools

XLIFF TRANSLATION EDITOR

TMX EDITOR

EXTR

Benutzerhandbuch Araya Bilinguale Termextraktion

BiLingual TermExtractor

Anleitung zum Einsatz des Terminologie-Extraktionswerkzeuges

Der bilinguale Extraktor

- Der bilinguale Extraktor ist ein einfach zu bedienendes und effizientes Werkzeug zum automatischen Generierung von Termpaaren aus übersetzten Dokumenten (TMX Dateien)
  - Ein Termpaar ist dabei eine Übersetzung aus Ausgangs- und Zielterm (Begriff)
  - Ein Term (Begriff) kann aus mehreren Wörtern bestehen.
- Diese Termpaare dienen z.B. zum Aufbau bzw. der Ergänzung der erarbeiteten Terminologie.

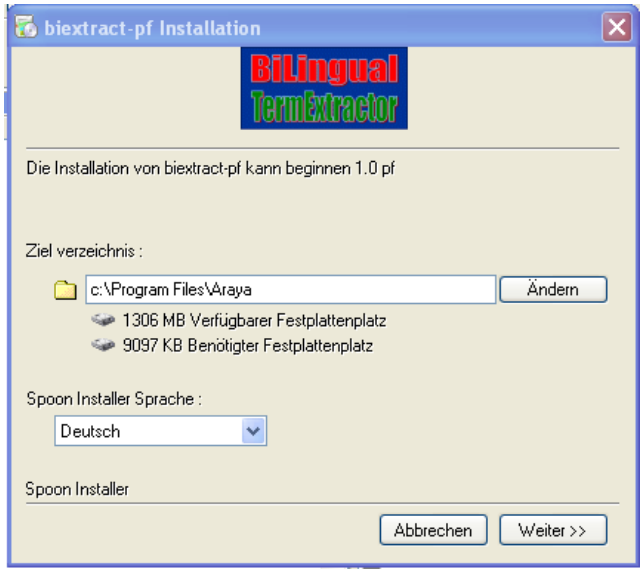
Versionen

- Der Extraktor wurde gemeinsam von der Heartsome Europe GmbH und LNE International entwickelt.
- Er ist als
  - Einzelplatzversion erhältlich.
  - er ist Teil der Araya Server Übersetzungswerkzeuge.

Kurzanleitung zum Extrahieren

- Extrahieren Sie die Begriffe
  - Datei -> Extrahiere Bilinguale Terminologie aus Datei
    - (Option: Öffne Extraktionsdatei nach Extraktion)
- Prüfen Sie die extrahierten Begriffe
  - Markieren Sie korrekte Übersetzungen als „validiert“
- Exportieren Sie die validierten Übersetzungen
  - Exportiere validierte Terme …

Installation



- Die Installation wird in das Verzeichnis **c:/Program Files/Araya** durchgeführt. Es wird empfohlen, diese Einstellung nicht zu ändern, da alle Initialisierungsdateien darauf ausgelegt sind.

Starten des Araya Extraktionswerkzeuges

- Gehen Sie zum Verzeichnis:
  c:/Program Files/Araya
  Starten: BiEdit.exe
- Oder Doppelklick auf:

Der Extraktionsansatz aus einer TMX Datei

- Aus einer TMX Datei werden mögliche Übersetzungspaare ermittelt. Dazu wird ein statistischer Ansatz verwendet, der die Häufigkeit des Auftretens von Termpaaren in der Ausgangs- und Zielsprache ermittelt.
- TMX = XML Austauschformat für Übersetzungsdatenbanken

Segment

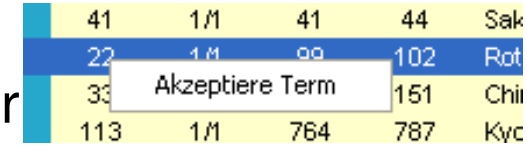
- Die Extraktion erfolgt auf der Basis von Segmenten, die in einer TMX Datei abgespeichert sind.
- Ein Segment kann dabei jeweils ein Satz oder ein Abschnitt sein.
- Formate in der TMX Datei werden ignoriert.

Bewertung und Validieren

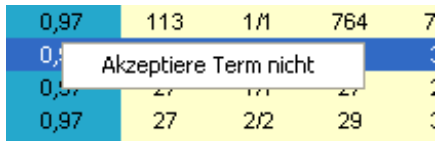
- Jedes gefundene Termpaar wird mit einem Qualitätswert versehen
  - 2. Spalte in der Extraktionstabelle
  - Wert liegt zwischen 1,0 (höchste Wahrscheinlichkeit, dass Paar zusammenpasst) bis 0,5 (geringste Wahrscheinlichkeit, dass Paar zusammenpasst)
- Terme können validiert werden, als zutreffend markiert werden
  - Letzte Spalte der Tabelle
    - Approved = geprüft = validiert
    - Unapproved = noch (nicht) validiert
- Validierte Werte können exportiert werden

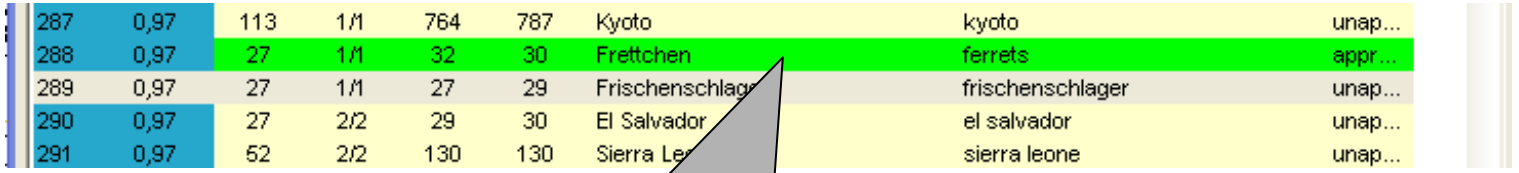
Validieren eines Termextraktionspaares

- Selektiere des zu validierenden Termpaares
- Validiere mit
  - Doppelten Mausklick auf Termpaar
  - Rechten Mausklick



- Entferne Validierungsmarkierung durch
  - Doppelten Mausklick auf Termpaar
  - Rechten Mausklick

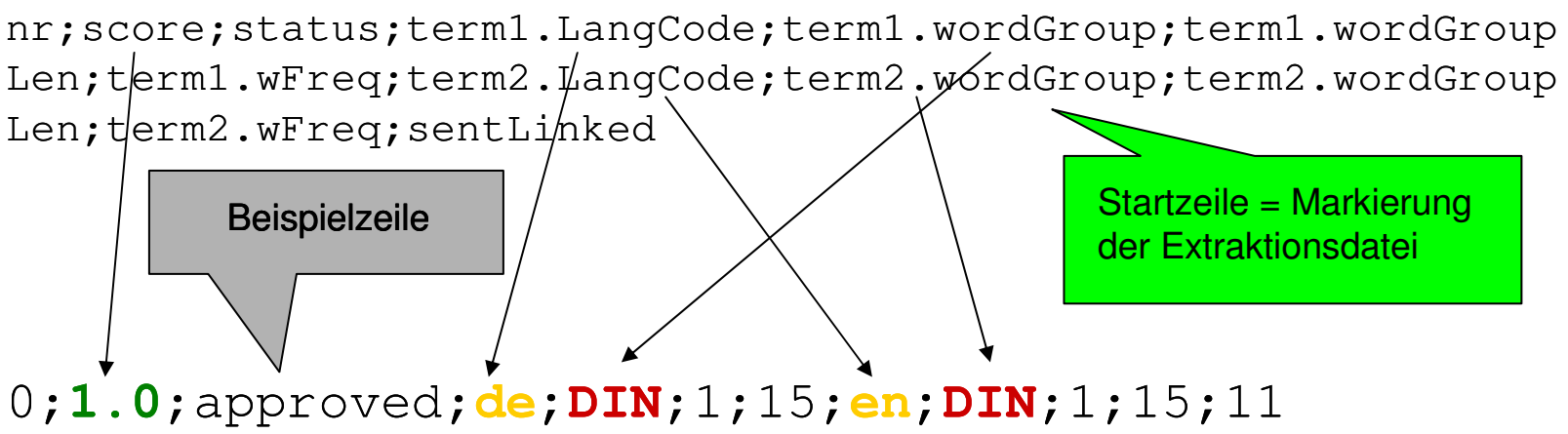




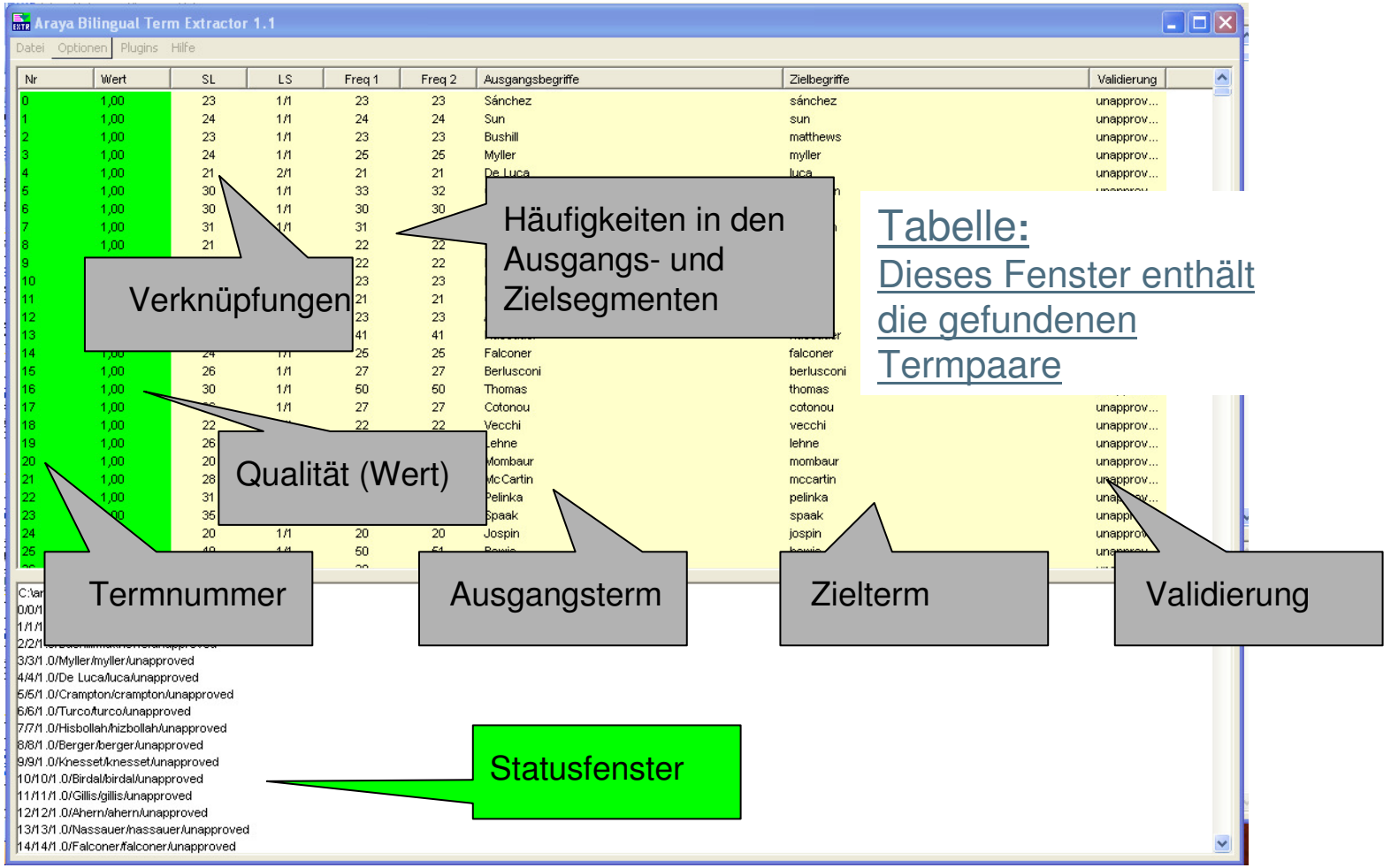
Validierte Terme werden grün angezeigt

Die Extraktionsdatei

- Eine Extraktionsdatei hat folgendes Format



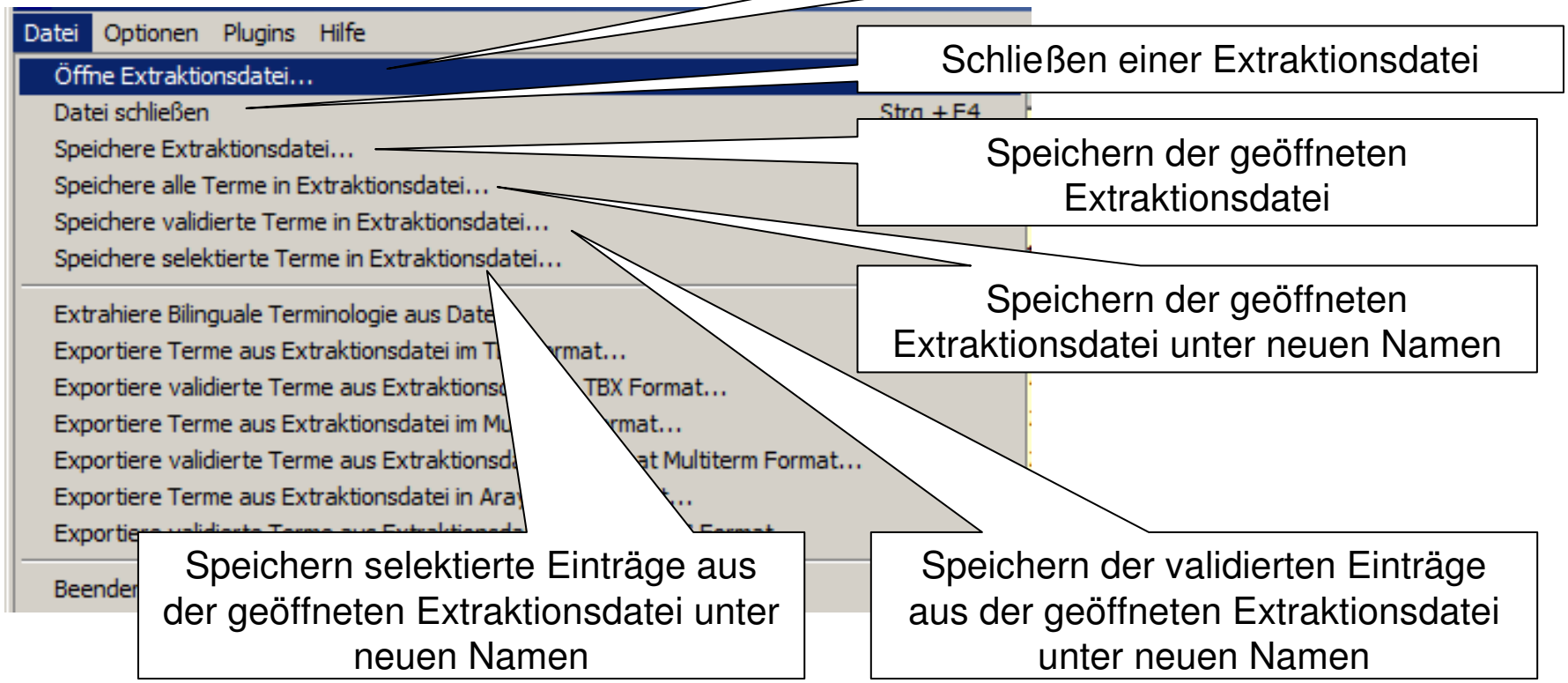
Die Extraktionsoberfläche

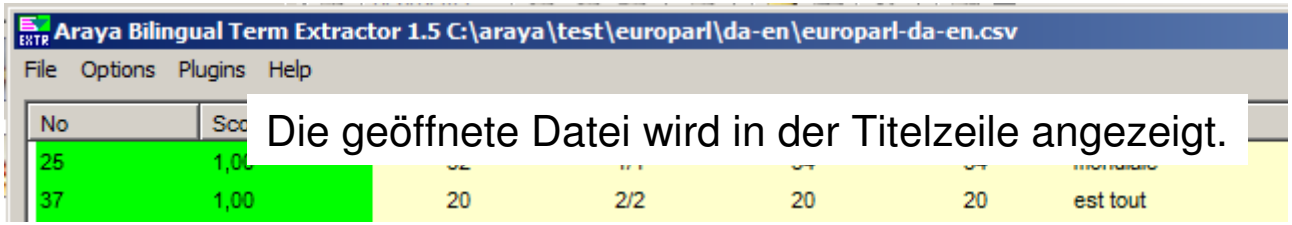


Die Spalten

- Wert
  - Statistisches Maß für die Wahrscheinlichkeit, dass Ausgangs- und Zielbegriff (Term) Übersetzungen sind; ein Qualitätsmaß
- SL
  - Die Anzahl der Satzpaare, in der sowohl der Ausgangs- als auch der Zielbegriff vorkommt.
- Freq 1
  - Anzahl der Sätze, in der der Ausgangsbegriff vorkommt
- Freq 2
  - Anzahl der Sätze, in der der Zielbegriff vorkommt
- Quellbegriffe
  - Der Ausgangsbegriff
- Zielbegriffe
  - Die Übersetzung des Ausgangsbegriffs
- Validierung
  - Auswahlbox, zur Markierung von korrekten Termpaaren

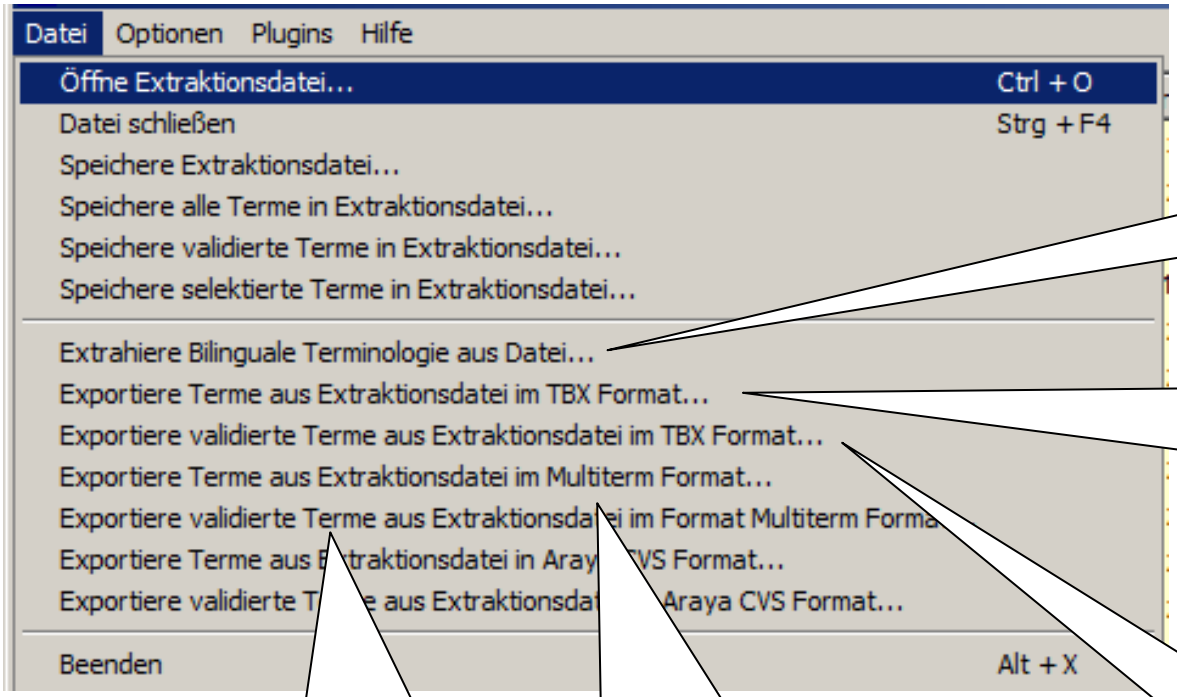
Das Datei Menü 1





Die geöffnete Datei wird in der Titelzeile angezeigt.

Das Datei Menü 2



Extrahieren der Termpaare aus einer TMX Datei

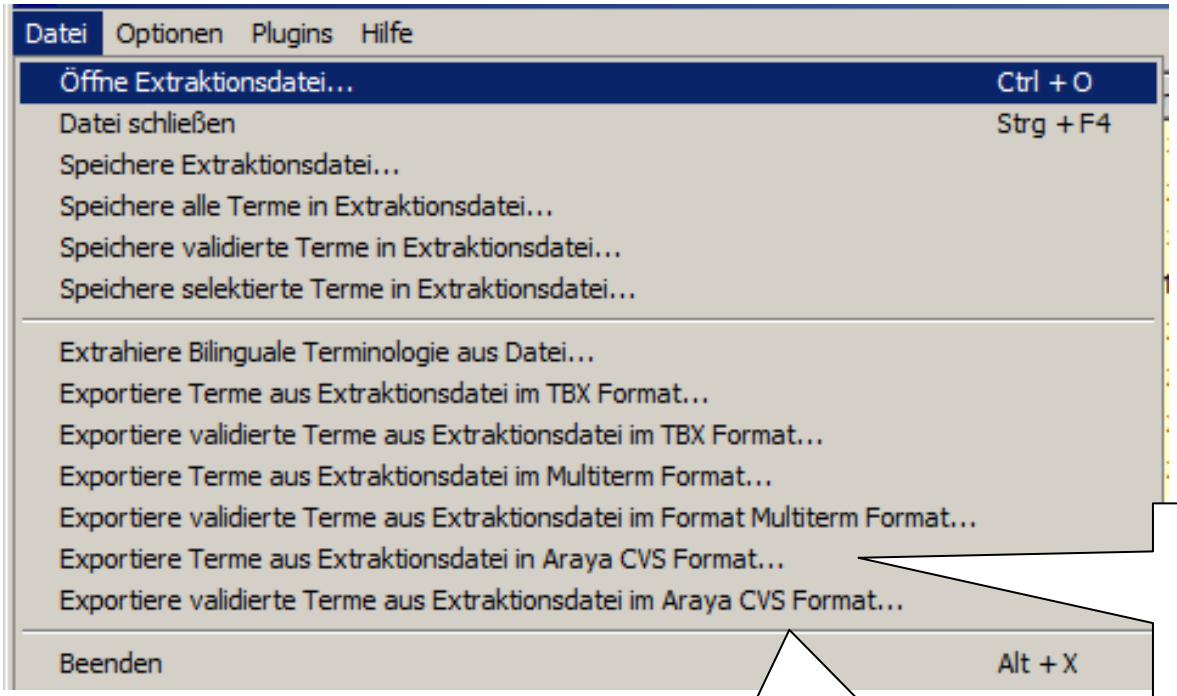
Exportieren der Einträge aus der geöffneten Extraktionsdatei im TBX Format

Exportieren der validierten Einträge aus der geöffneten Extraktionsdatei im Multitermformat

Exportieren der Einträge aus der geöffneten Extraktionsdatei im Multitermformat

Exportieren der validierten Einträge aus der geöffneten Extraktionsdatei im TBX Format

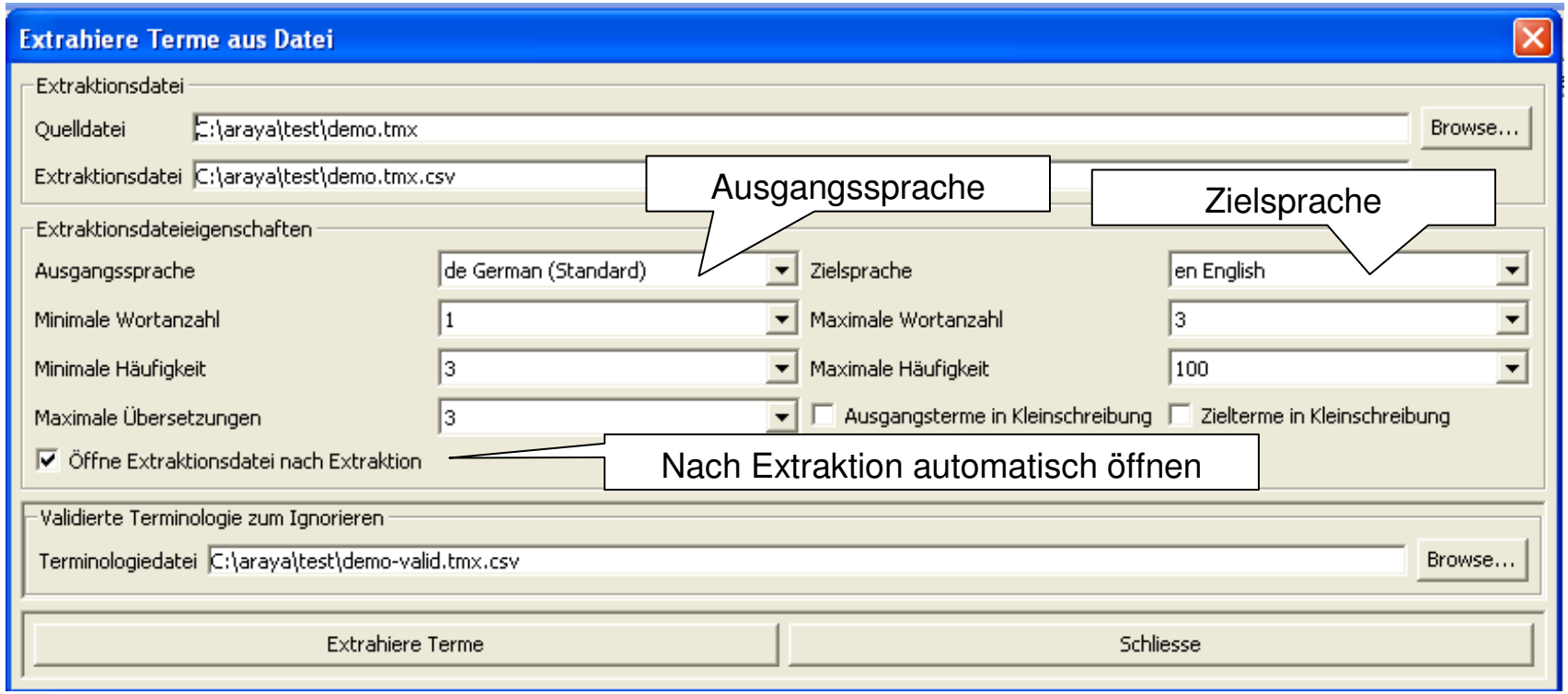
Das Datei Menü 3



Exportieren der Einträge aus der geöffneten Extraktionsdatei im Araya CSV Format

Exportieren der validierten Einträge aus der geöffneten Extraktionsdatei im Araya CSV Format

Extrahieren der Termpaare aus einer TMX Datei



Extraktionsparameter 1

- Minimale / Maximale Wortanzahl
  - Damit wird gesteuert, wie viele Worte mindestens und höchstens im Begriff enthalten sein sollen
- Minimale / Maximale Häufigkeit
  - Damit wird gesteuert, wie oft der extrahierte Begriff mindestens und höchstens vorkommen darf
- Maximale Übersetzungen
  - Damit wird gesteuert, wie viele Übersetzungen maximal gefunden werden sollen
- Ausgangs/Zielterme in Kleinschreibung
  - Damit wird gesteuert, ob die Worte der extrahierten Begriffe in Kleinschreibung umgewandelt werden sollen

Extraktionsparameter 2

- Validierte Terminologie zum Ignorieren
  - Wenn hier eine Extraktionsterminologiedatei angegeben wird, werden bei Extrahieren alle Terme, die in dieser Datei als „validiert“ gekennzeichnet sind, ignoriert.
  - Damit werden schon bekannte Übersetzungen ignoriert.
- Nach dem Start der Extraktion wird ein Statusfenster angezeigt.

Exportieren

- Beim Exportieren werden die Einträge der geladenen Extraktionsdatei in verschiedene Format geschrieben.
  - TBX
    - Name der Extraktionsdatei + „.tbx“
  - Multiterm (™ of Trados/SDL International)
    - Name der Extraktionsdatei + „.multiterm“
  - Araya CSV
    - Name der Extraktionsdatei + „araya.csv“
  - Zeichencodierung ist dabei immer UTF-8
- Es können dabei entweder alle Einträge oder nur die validierten Einträge geschrieben werden
- Zusätzlich dient der im „Optionen-Menü“ eingestellte Qualitätswert (Export Wertefilter) als Selektionskriterium.
  - Je nach eingestelltem Wert werden nur die Wert mit mindestens der jeweiligen Höhe (z.B. > 0.6) exportiert.

Araya CSV Format

- Das Araya CSV Format enthält in der ersten Zeile das Sprachpaar gefolgt von den extrahierten Begriffen

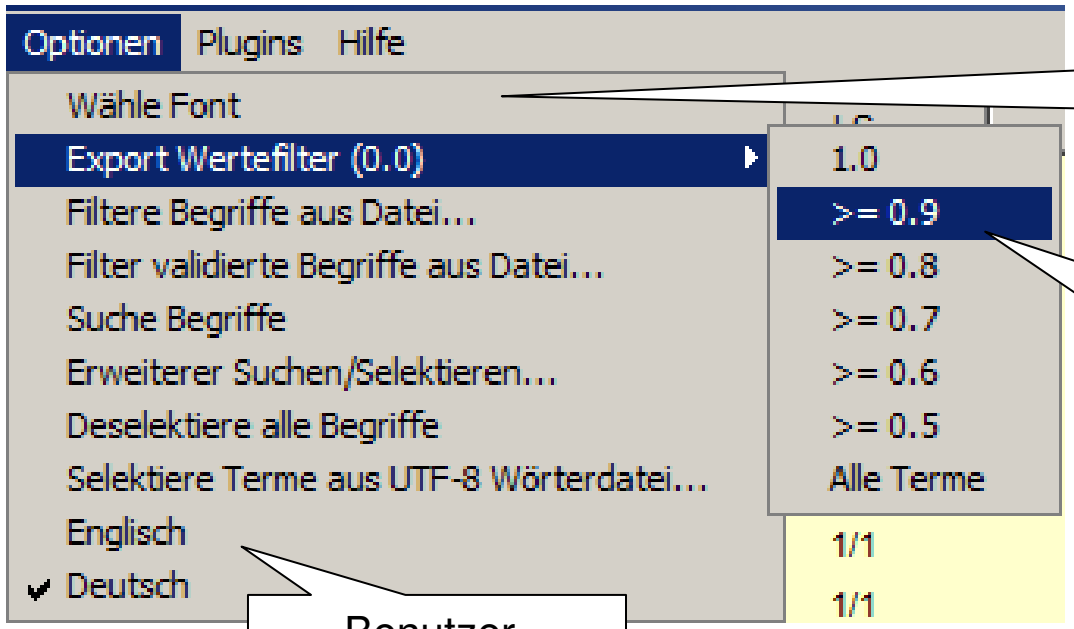
Beispiel

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

Sprachen durch ; getrennt

Extrahierte Terme durch ; getrennt

Das Optionen Menu 1

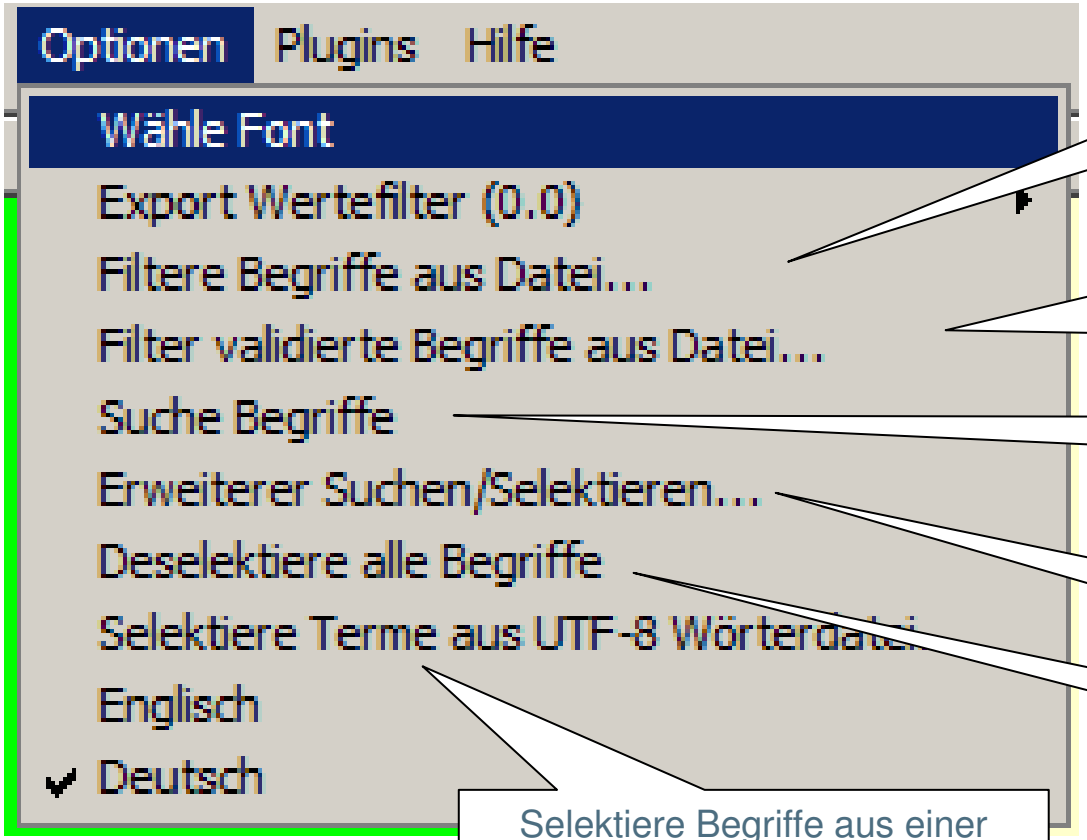


Zeichensatz für Anzeigefenster und Tabelle

Einstellungen der Minimalqualität für den Export der Termpaare

Benutzer-schnittstellen-sprache

Das Optionen Menu 2



Entferne alle Begriffe, die in einer Termextraktionsdatei gefunden werden.

Entferne alle Begriffe, die in einer Termextraktionsdatei gefunden werden und validiert sind.

Suche Begriffe in Termtabelle

Erweiterte Suche in Quell- und Zielbegriffen

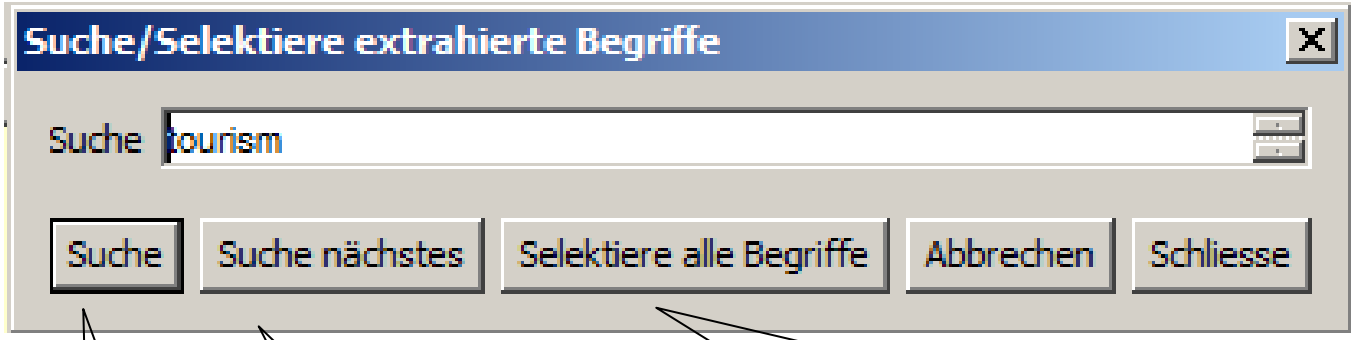
Deselektiere alle Terme

Selektiere Begriffe aus einer Wortliste in einer Datei

Filter Term Funktionen

- Die Filterfunktionen entfernen alle Terme aus der Termtabelle, die in einer anderen Termextraktionsdatei enthalten sind.
- Die identischen Terme werden entfernt.
- Je nach gewählter Methode betrifft diese die validierten oder alle Termen in der gewählten Termextraktionsdatei.

Suche Begriffe

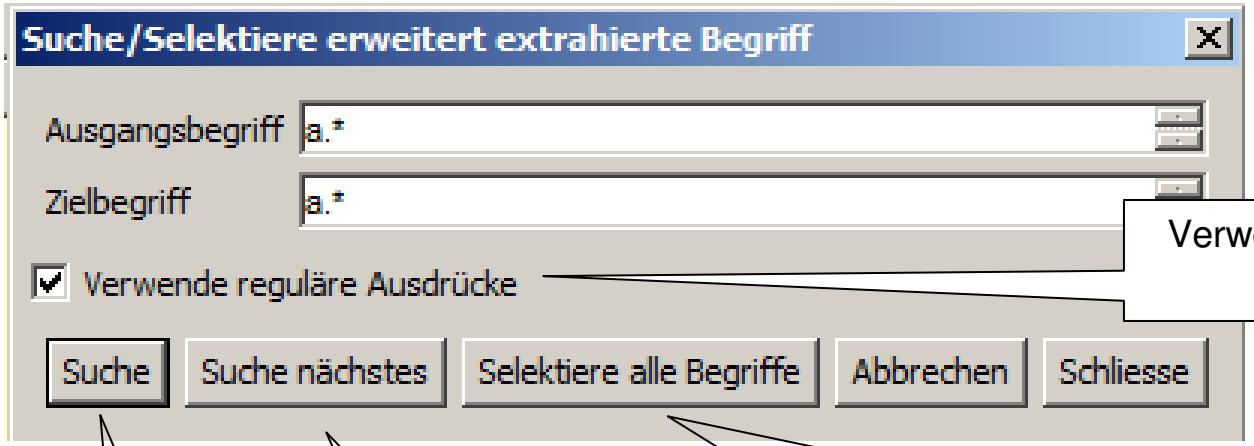


Suche mit diesem Begriff und selektiere alle passenden Einträge in der Tabelle. Die selektierten Einträge können mit "Datei -> Speichere selektierte Einträge in Extraktionsdatei…" gesichert werden.

Suche nächsten passenden Begriff

Starte Suche mit diesem Begriff

Erweiterte Suchfunktionen



Verwende reguläre Ausdrücke zur Suche

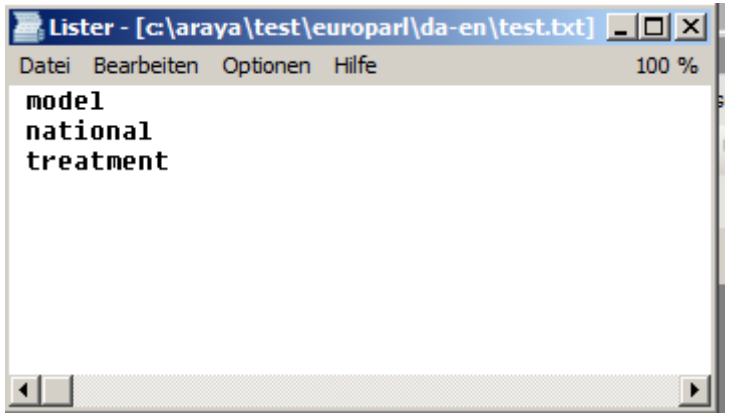
Suche mit diesem Begriff und selektiere alle passenden Einträge in der Tabelle. Die selektierten Einträge können mit "Datei -> Speichere selektierte Einträge in Extraktionsdatei…" gesichert werden.

Suche nächsten passenden Begriff

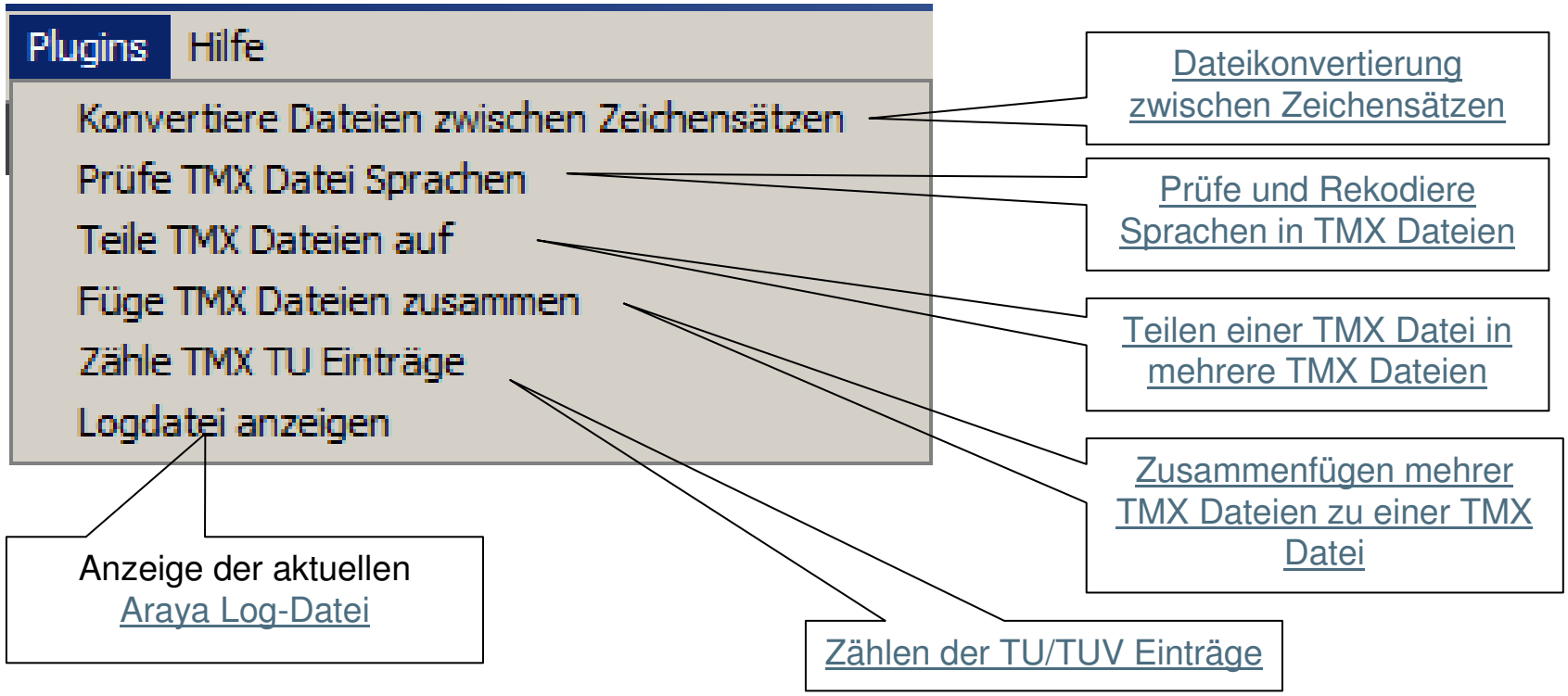
Starte Suche mit diesem Begriff

Selektiere Terme aus UTF-8 Datei

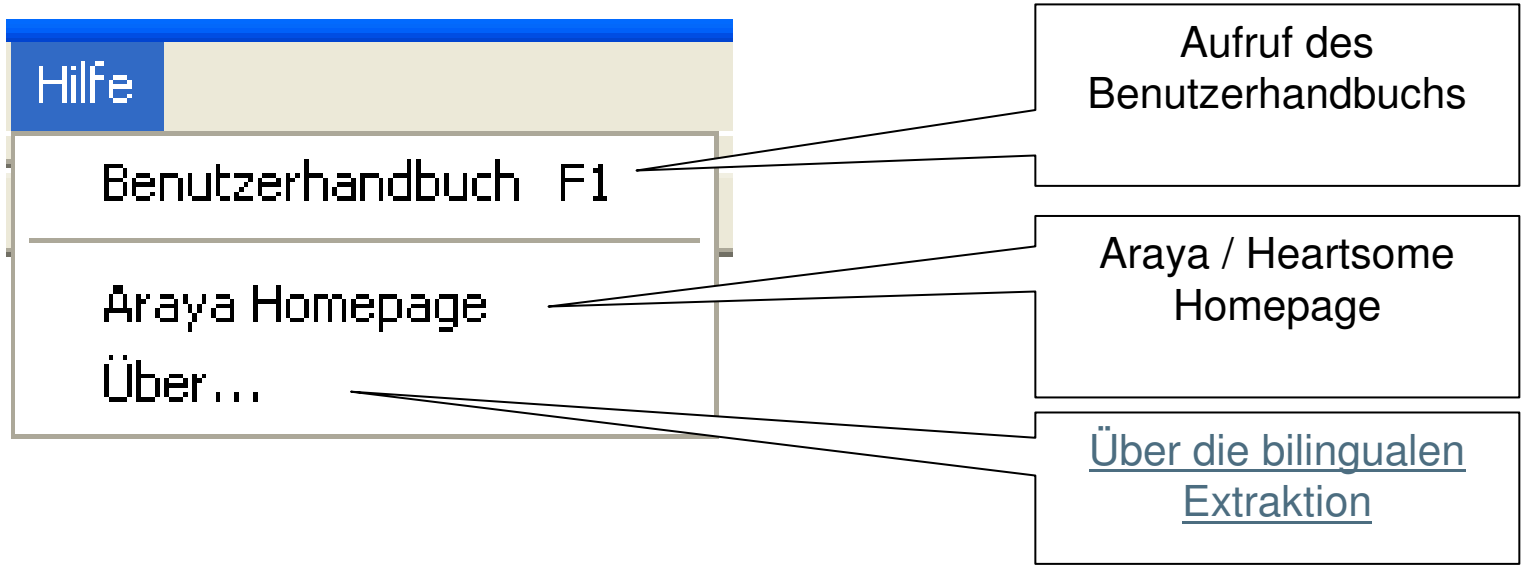
- Dieser Filter selektiert alle Terme in der Tabelle, die in einer UTF-8 kodierten Textdatei enthalten sind. Die Filtersuche sucht sowohl im Quell- als auch Zielbegriff.
  - Es selektiert auch Teilzeichenketten.
- Die selektierten Einträge können mit “Datei -> Speichere selektierte Einträge in Extraktionsdatei…” gesichert werden.
- Jedes Wort (Zeichenkette) in der Datei muss in einer Zeile vorkommen.



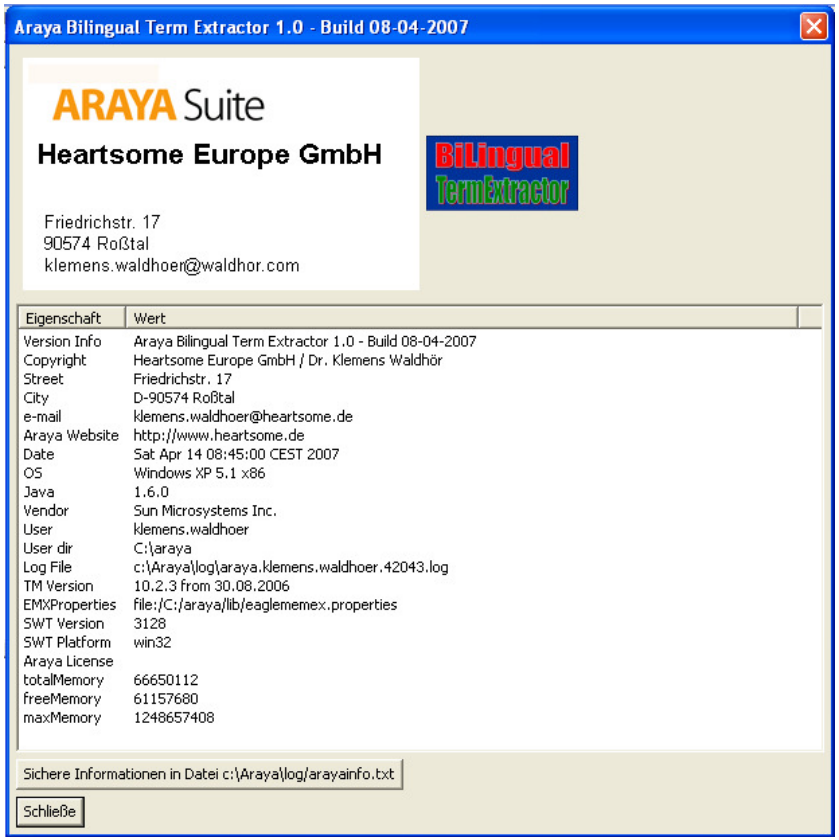
Das Plugins Menü



Das Hilfe Menü

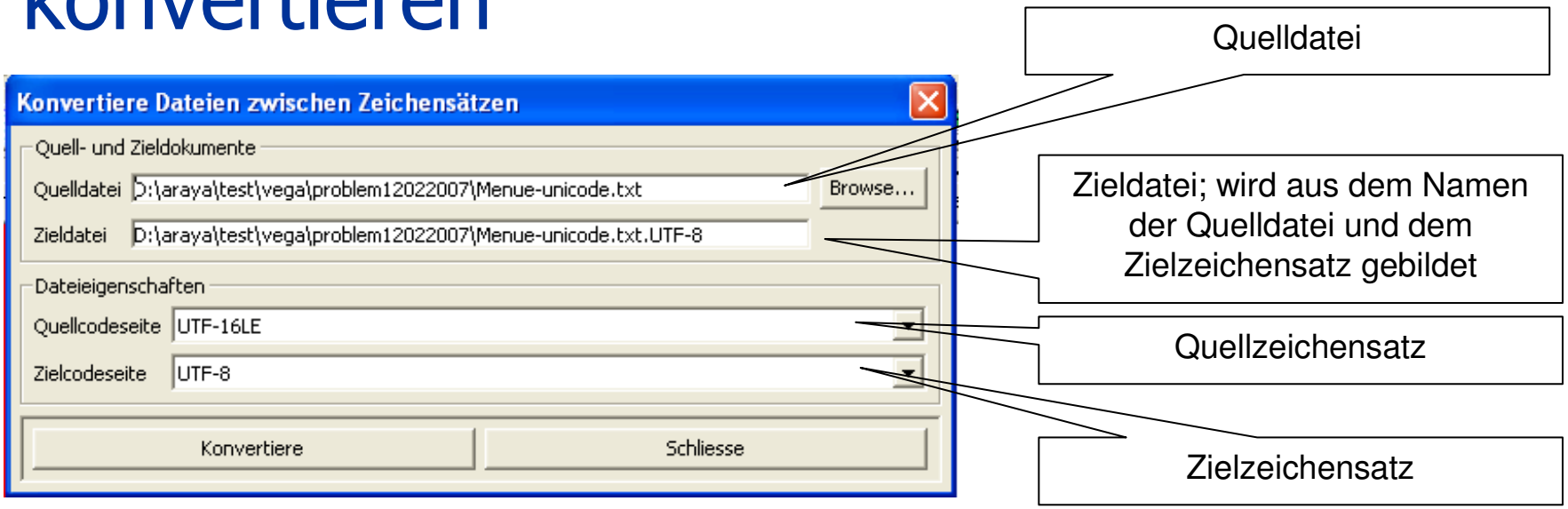


Über die Extraktion



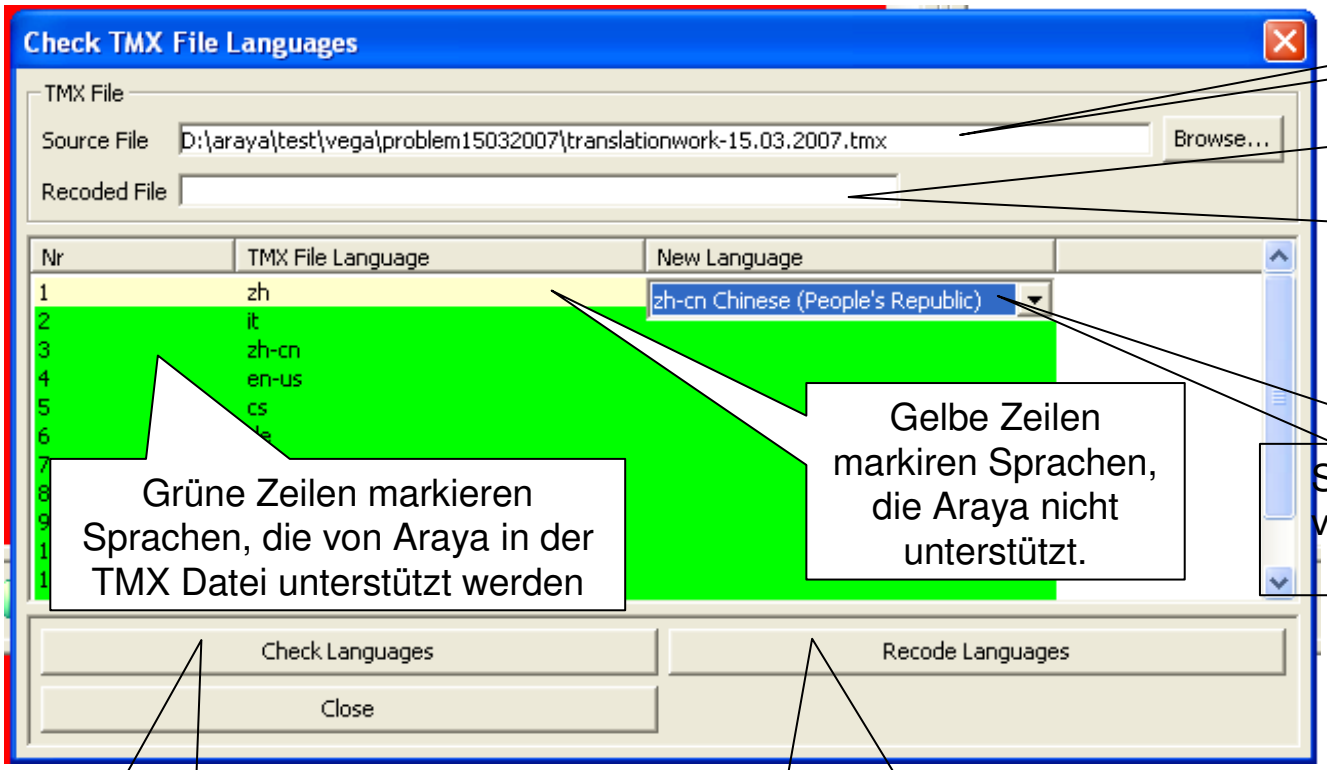
- Hier werden Basisinformationen über die Extraktions-Konfiguration angezeigt.
- Dies Informationen könne für eine mögliche Fehlerdiagnose wichtig sein
- Mit „Sichere…“ kann diese Information der angegebenen Datei gespeichert werden.

Dateien zwischen Zeichensätzen konvertieren



Diese Funktion erlaubt es Dateien von einem Zeichensatz in einen anderen zu konvertieren. Je nach Zielzeichensatz können auch zwei Ausgabedateien geschrieben werden. Wenn es sich um eine UTF-8, 16, 32 oder UCS Datei handelt, werden zwei Dateien geschrieben. Die zweite Datei, mit der zusätzlichen Erweiterung „.**nobom**“, ist eine Kopie der ersten, nur werden aus ihr die BOM (Byte Order Marker) entfernt. Diese Datei zum Importieren zu verwenden empfiehlt sich insbesondere bei UTF-8 Import-Dateien für Araya, da die Java Lese-Funktionen für UTF-8 Dateien die BOMs nicht entfernt und diese beim Einlesen als normale Zeichen eingelesen werden würden (und damit zu fehlerhaften Einträgen führen würden). (Dies ist eine bekannter Fehler von Java, wird aber von SUN nicht behoben!).

Prüfen/Rekodieren von TMX Dateien



TMX Dateiname

Rekodierte TMX Dateiname; Besteht aus dem TMX Dateinamen ohne „.tmx“ Extension und neuer Endung „.recoded.tmx“

Grüne Zeilen markieren Sprachen, die von Araya in der TMX Datei unterstützt werden

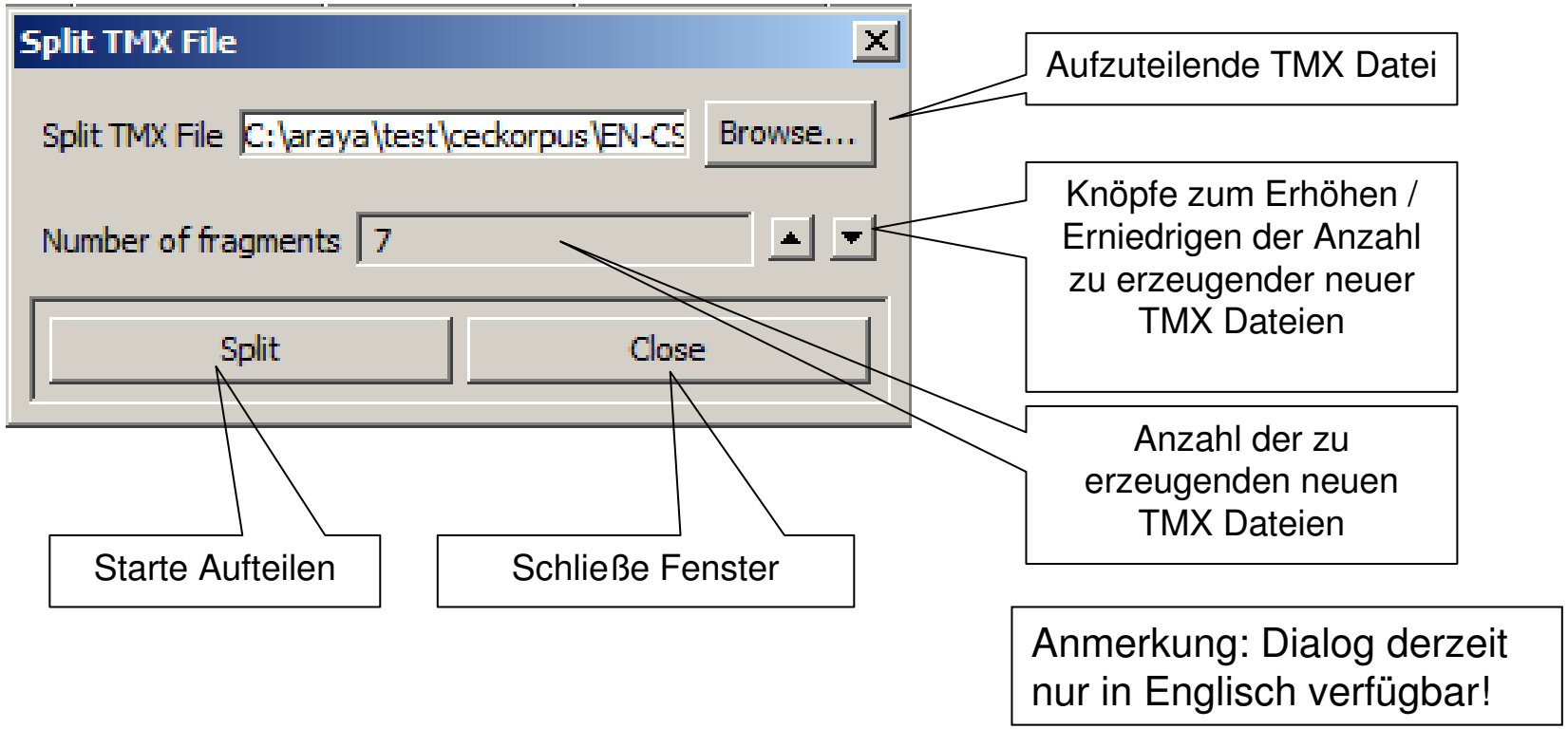
Gelbe Zeilen markiren Sprachen, die Araya nicht unterstützt.

Sprachkodierung, die von Araya unterstützt werden

Starte Prüfung der Sprachen in TMX Datei

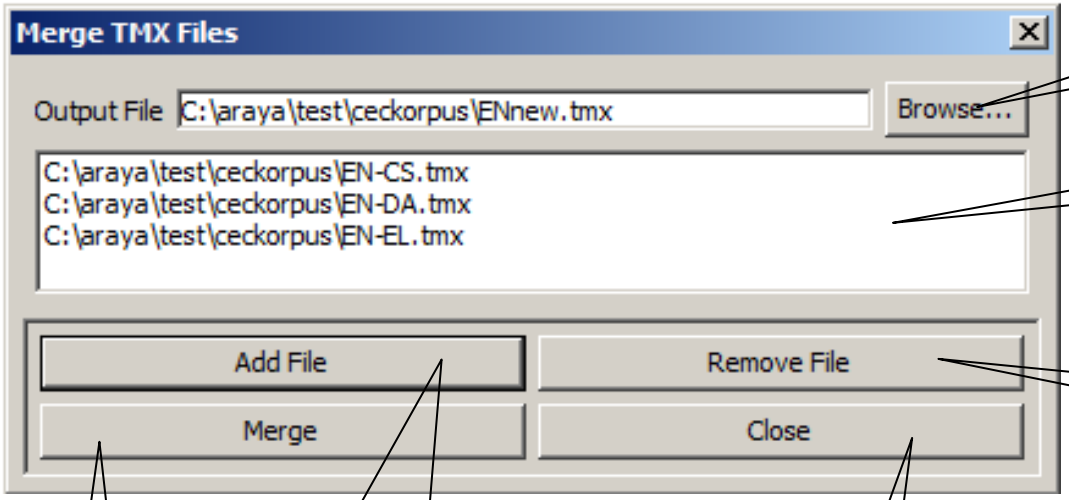
Rekodiere Sprache in der TMX Datei basierend auf den gewählten neuen Sprachcodes

Teilen einer TMX Datei



Die neu erzeugten TMX Dateien werden aus dem Namen der alten Datei und der jeweiligen Zahl von 1 bis n zusammengesetzt.

Zusammenfügen von TMX Dateien



Wähle Name für neue TMX Datei

Liste gewählter TMX Dateien

Entferne Datei aus der Liste

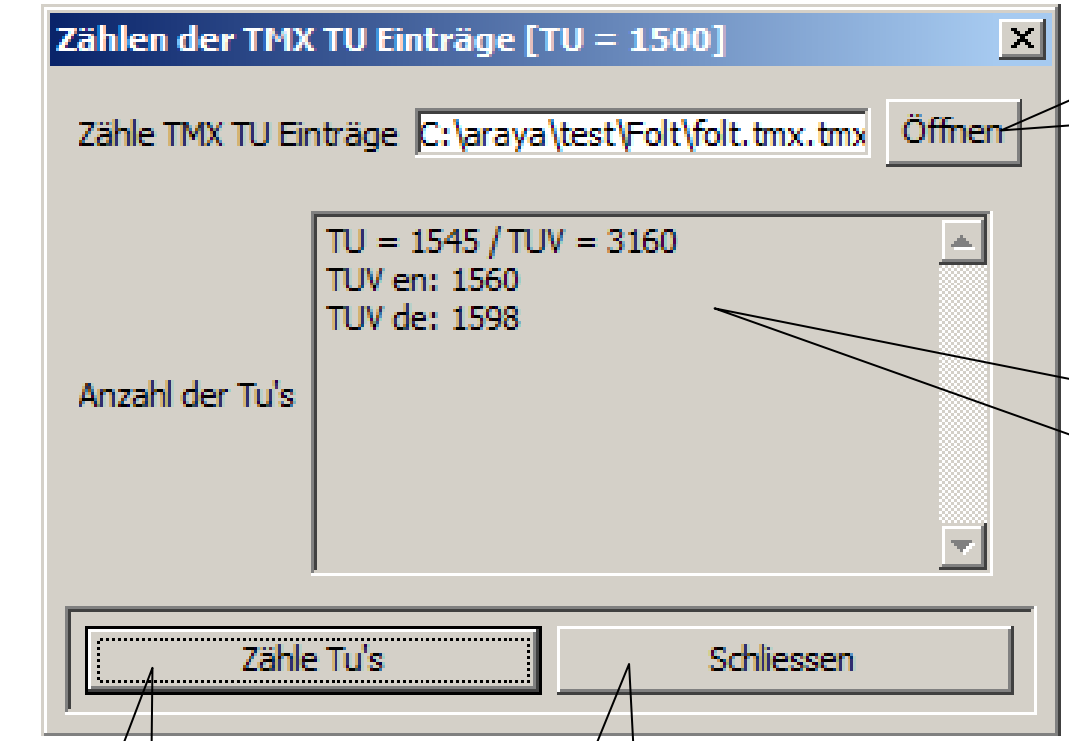
TMX Datei zur Liste hinzufügen

Anmerkung: Dialog derzeit nur in Englisch verfügbar!

Starte Zusammenfügen

Schließe Fenster

Zählen von TUs/TUVs in TMX Datei



Auswählen der TMX Datei zum Analysieren

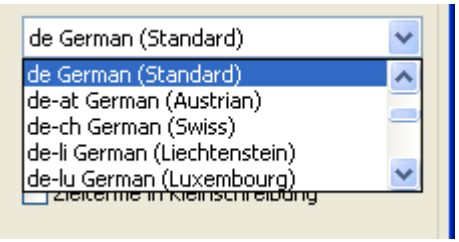
Resultat:
Erste Zeile: Anzahl TUs und TUVs
Folge Zeile:
Anzahl sprachspezifischer TUVs

Starte Zählung

Schließe Fenster

Hinzufügen von Sprachencodes

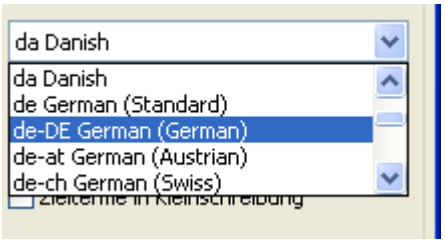
- Vordefinierte Sprachencodes sind in der Datei file „ini/lancodes.txt“ definiert.
- Weitere Sprachencodes können durch erweitern dieser Datei hinzugefügt werden.
- Ein Beispiel:



Hinzufügen de-DE: Eine Zeile wie hinzufügen, wobei = den angezeigten Namen der Sprache und Sprachecode trennt.

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



Impressum

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

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