

Workshop

Peter Sandrini

University of Innsbruck, Austria

Control over Digital Technology Free and Open Source CAT Tools

Timisoara

March 26, 2015

Abstract

In a digitalized and globalized world, translation technology is becoming an inevitable part of translation. It not only concerns translators but also users of translation, trainers of translators, and localizers. Translation Technology can boost the efficiency and consistency of translation, but inconsiderate use of software and services may also cause translators losing control over the translation process and translation data.

The workshop outlines the concept of translation technology as well as free and open source software and presents two compilations of available free translation technology tools developed at the University of Innsbruck: USBTrans – a collection of preinstalled packages on a USB stick, and tuxtrans – a tailor-made Linux distribution for translators.



Contents

- 1) Control over digital technology
Free Software
Free translation technology packages:
USBTrans and tuxtrans
- 2) Typical work tasks:
 - ✓translate a website
 - ✓create a TM on the basis of existing translations
 - ✓manage terminology and dictionaries
 - ✓extract terminology from texts
 - ✓use machine translation
 - ✓convert file formats
 - ✓manage bilingual files
 - ✓manage pdf files
 - ✓quality assurance
 - ✓use text corpora



Dominant Technology?

- „it is hard to think of a business process that is not wholly, or partly, dependent on technology“
- what about language services and translation?
„technological turn in translation“ (Cronin 2010)
- costs and risks of technology
- translation technology $\neq \geq$



Translation Technology

- any kind of digital Information and communication technology (ICT)
- which supports or performs the translation process
- with the aim of meeting adequate efficiency and quality requirements



Control

- choice vs (being) use(d) (independence)
- confidentiality
- integrity of program code
- integrity of data
- availability of data
- overall management of the IT
- configuration, installation of software



Control: examples

Choose your software independently
do not let translation agencies dictate your choice

Do not let financial factors limit your choice

Do not let licenses limit your freedom
I want to install my software on a desktop and
on a notebook computer as well as on my network
I want hassle-free updates

Be social and share your software with friends
the easiest way to guarantee cooperation
and data exchange with colleagues

Be confident about your data
I have a confidentiality agreement with my clients
my translation data are economic assets



Win back and maintain control



**FREE SOFTWARE
GIVES YOU BACK
CONTROL OVER
YOUR COMPUTER**

Free and Open-Source

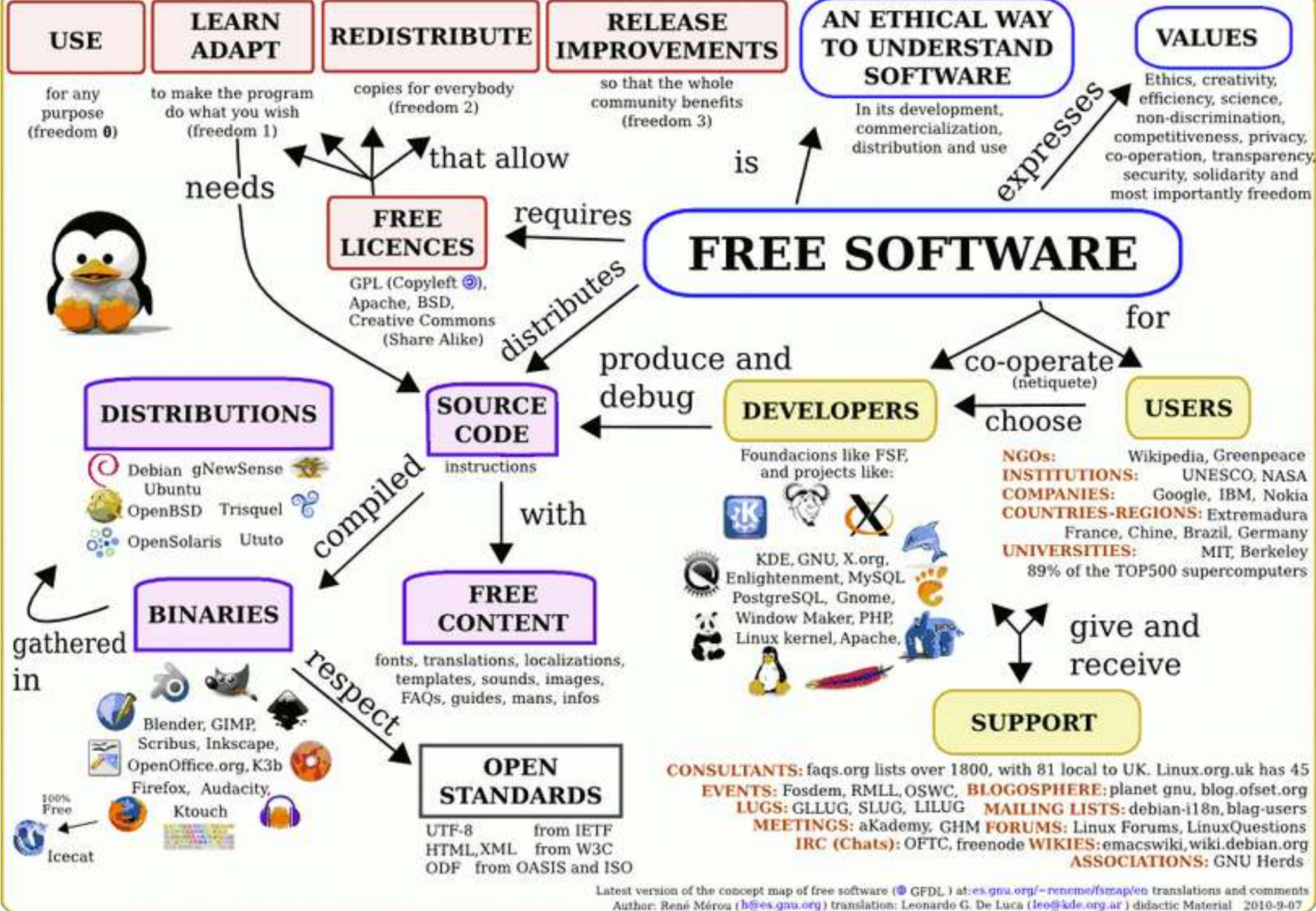
Freedom to

- run the program as you wish, for any purpose (freedom 0).
- study how the program works, and change it so it does your computing as you wish (freedom 1). Access to the source code is a precondition for this.
- redistribute copies so you can help your neighbor (freedom 2).
- distribute copies of your modified versions to others (freedom 3). By doing this you can give the whole community a chance to benefit from your changes. Access to the source code is a precondition for this.



Licenses:

- GNU GPL
- Apache License 2.0
- BSD 2/3
- (L)PGL
- MIT license
- Mozilla Public License 2.0
- Eclipse Public License
- Creative Commons



Latest version of the concept map of free software (© GFDL) at: es.gnu.org/~reneme/fsmap/en translations and comments
 Author: René Mérou (h@es.gnu.org) translation: Leonardo G. De Luca (leo@kde.org.ar) didactic Material 2010-9-07

Why use free software?

- allow a cost-saving start of your career
- facilitate cooperation with colleagues
- avoid copyright infringements
- full control over your own PC
- ease of use without a license or activation code
- participation in developing applications through online communities
- changes (dependent) consumers into (autonomous) agents

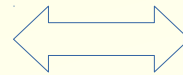


Structural differences

proprietary software

Translation Environment
Tools (TenT)
all-in-one application for
translators

- ✓ Translation-Memory
- ✓ Terminology-Management
- ✓ Alignment
- ✓ Search for collocations
- ✓ Analysis and statistics
- ✓ Project management
- ✓ Code-Protection
- ✓ Batch-scripts
- ✓ Spellchecking
- ✓ Code page conversion
- ✓ Format conversion
- ✓ ...



free software

individual projects with
specific functionality

- ✓ translation memory
- ✓ analysis and statistics
- ✓ code protection

- ✓ terminology management

- ✓ project management

- ✓ code page conversion
- ✓ format conversion

- ✓ Spell checking

- ✓ ...

TEnTs

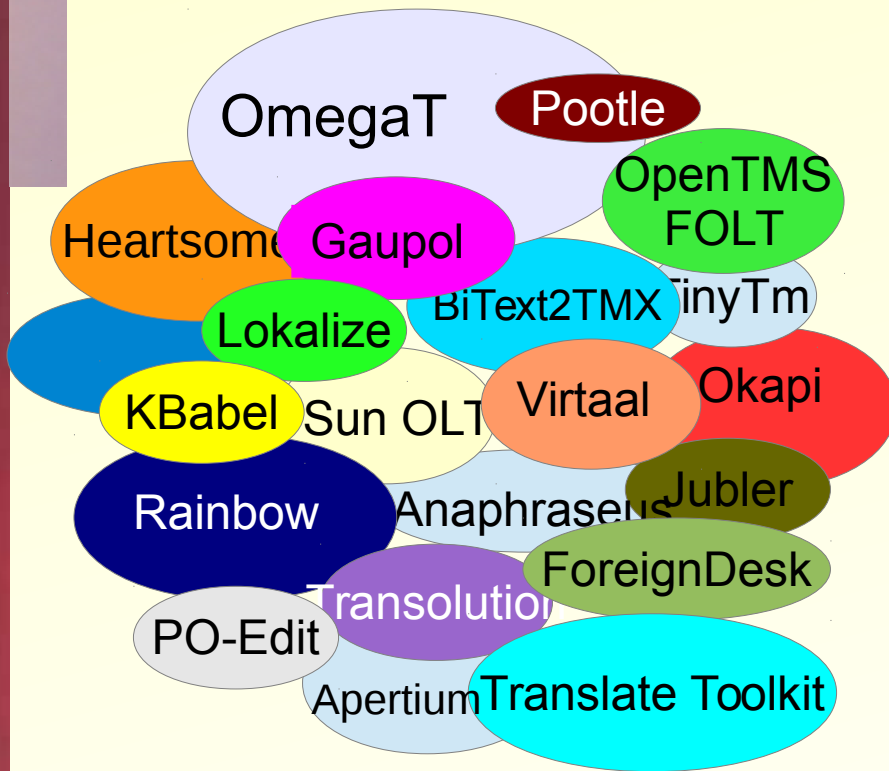
- Translation Environment Tools (TenT) all-encompassing application for translators
- Generic term for „Translation-Memory-System“ or „Computer aided translation CAT-System“



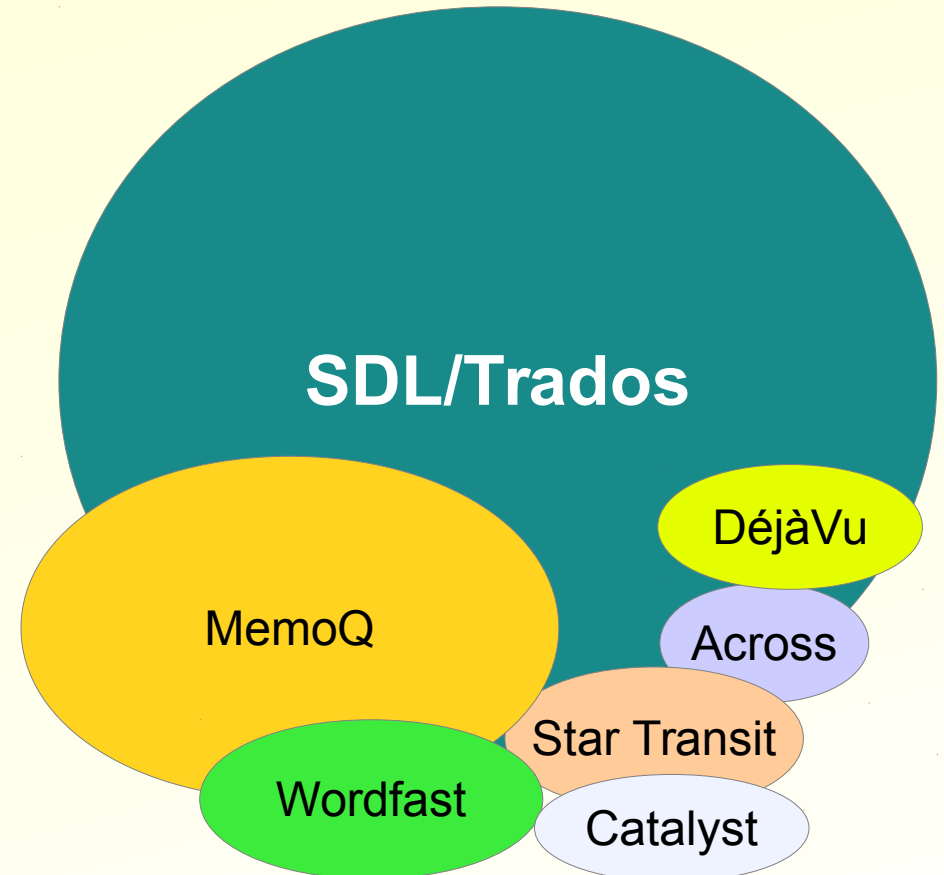
- ✓ Translation-Memory
- ✓ Terminology-Management
- ✓ Alignment
- ✓ Search for collocations
- ✓ Analysis and statistics
- ✓ Project management
- ✓ Code-Protection
- ✓ Batch-scripts
- ✓ Spellchecking
- ✓ Code page conversion
- ✓ Format conversion
- ✓ ...

market reality

open



proprietary



FOSTT compiled

1. USBTrans – for Windows

<http://homepage.uibk.ac.at/~c61302/fsftrans.html>



2. *tuxtrans* – Open Translation Desktop System

<http://www.tuxtrans.org>



USBTrans

- a compilation of translation related free and open source software which can be started from an USB stick without installation (Portable Apps)
- download from <http://homepage.uibk.ac.at/~c61302/fsftrans.html> some samples on your USB
- unpack the zip archive on your local pc and you are ready to go; you may also copy the unpacked files onto an USB stick (4GB) and start the programs from there
- just plug in the USB stick and start the USBTrans menu



tuxtrans



- complete desktop system for translators
 - based on Linux as a free operating system
 - and many specific application for translators
- multilingual
 - Italian, English, Spanish, German by default
 - with many more languages available online
- all open source or free software
- website: <http://www.tuxtrans.org>
Twitter: <https://twitter.com/tuxtrans>
- live-system, it can be started from your USB stick without installation



Requirements

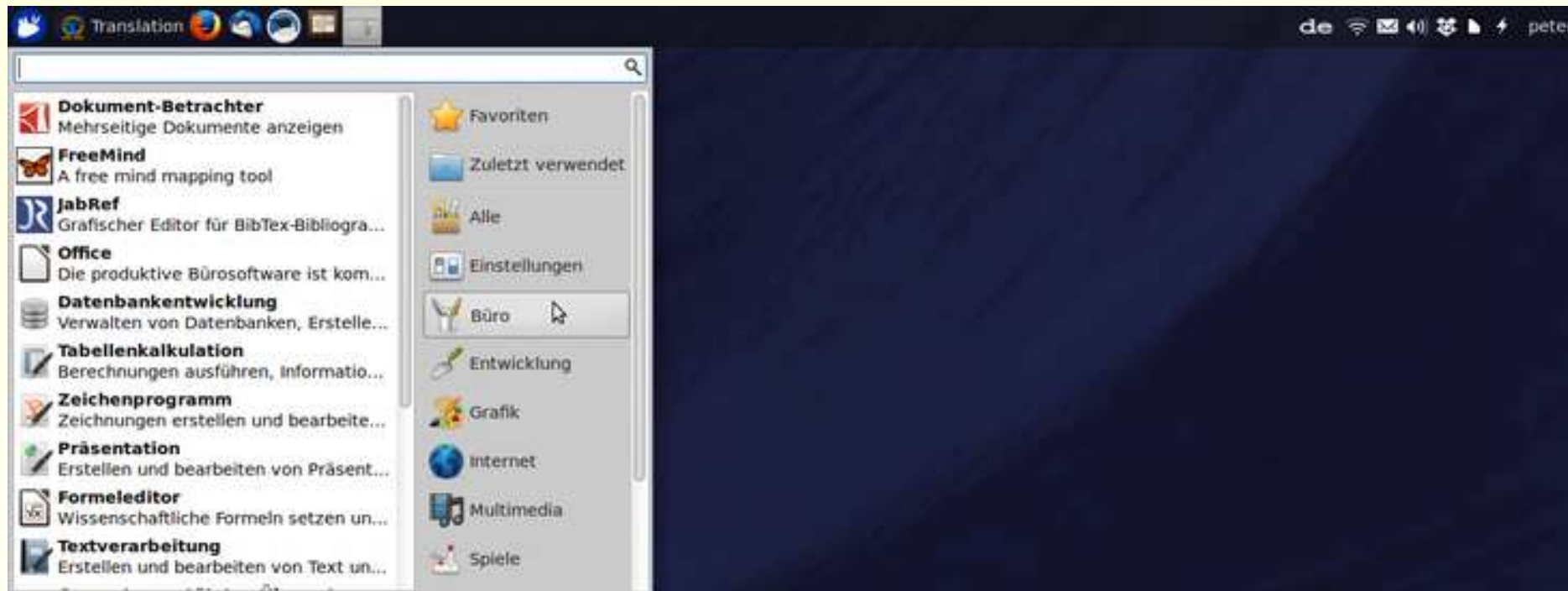
- operating system
- standard applications
- specific applications



=



**free
open-source**



Requirements

- operating system
- standard applications
- specific applications



=

**free
open-source**

Ask me about
**Free
Software**



why?

- to be able to perform all tasks related to translation
- to be able to use and install it ad libitum
- to be able to distribute it to translators and students

tuxtrans may be used as



- 1) Live-DVD
(without installation, slow performance)
- 2) Live-USB
(without installation, rather slow performance)
- 3) Virtual machine (VirtualBox VMWare)
- 4) second OS
(with installation, fast performance, easy to customize and adapt)
- 5) main operating system



prepare to migrate

- use cross-platform-applications
- and standard formats
odt, pdf, tmx, tbx, xcliff ...



standard apps



	<i>MS-Windows</i>	<i>Linux</i>
› word processors	LO-Writer	LO-Writer
› spreadsheets	LO-Calc	LO-Calc
› presentations	LO-Impress	LO-Impress
› databases	Access	MYSQL
› DTP	Xpress	Scribus
› image manipulation	Gimp	Gimp
› office suite	Libre Office	Libre Office
› Browser	Firefox	Firefox
› E-Mail	Thunderbird	Thunderbird

translate with FOSS

- common tasks of a free-lance translator
- with free and open source translation technology
- on a free digital infrastructure
- with *tuxtrans* the Linux Desktop for translators

tuxtrans
14.04



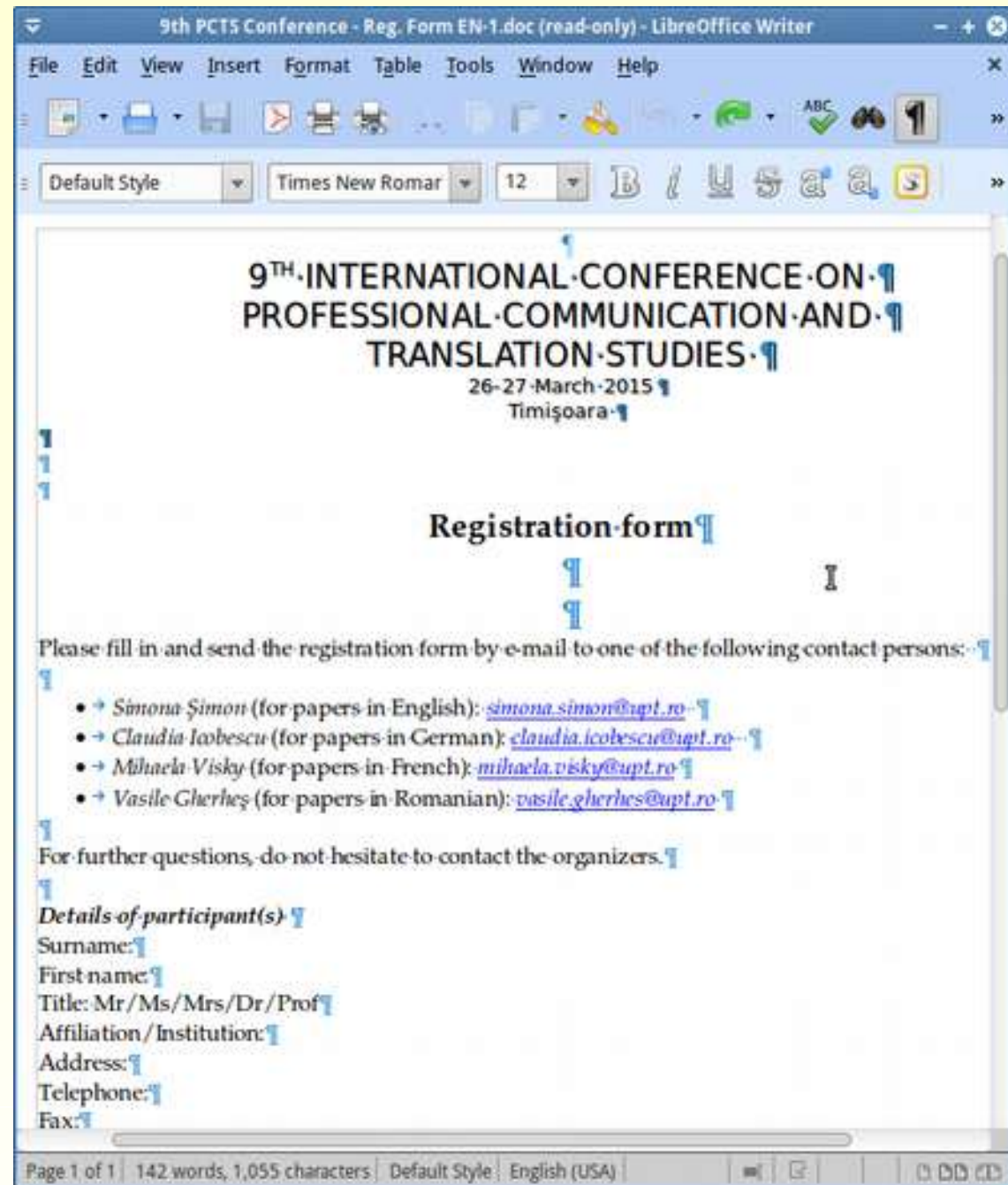
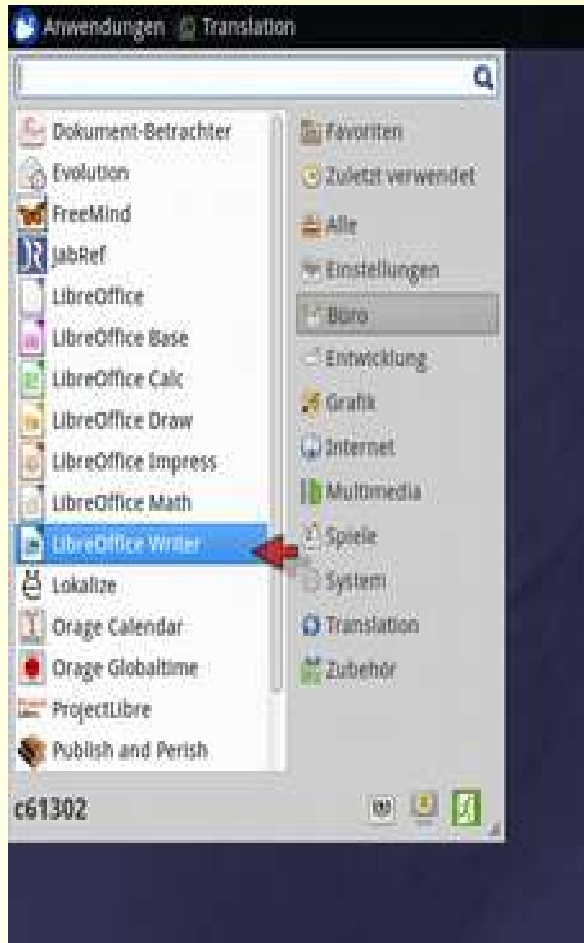
▶ translate a Word document with TM support

what you need:

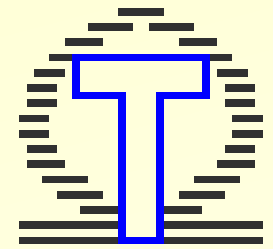
- a word processor
- a TM system
- translation memories
- terminology lists



sample text



OmegaT: supported formats



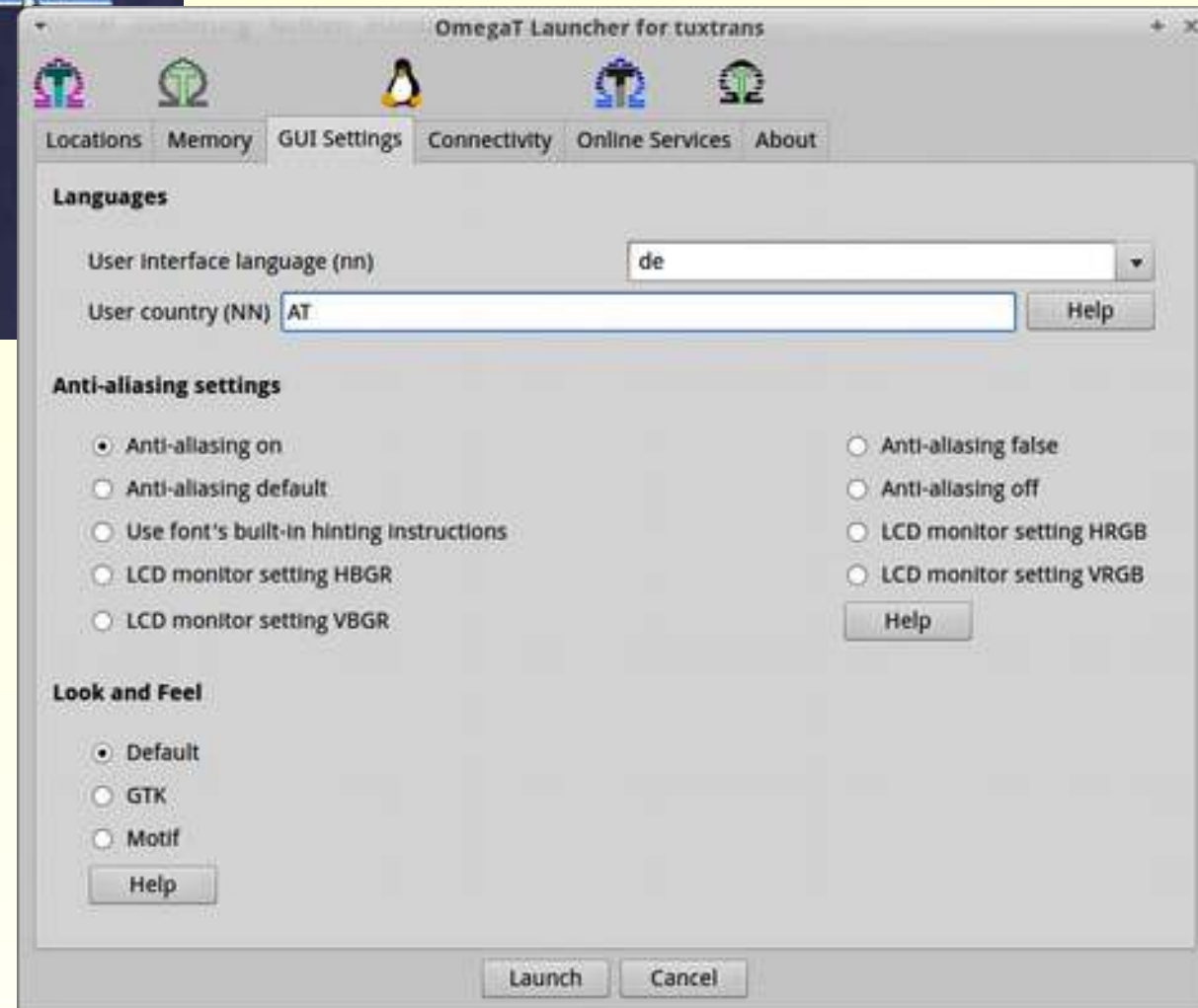
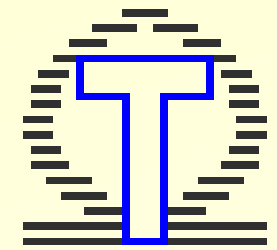
text formats

- Plain text (any encoding supported by Java), including Unicode
- StarOffice, OpenOffice.org, LibreOffice and OpenDocument
- **Open XML (Microsoft 2007/2010/2013)**
- (X)HTML (including complete website tree structure)
- Help & Manual
- HTML Help Compiler
- LaTeX
- DokuWiki
- CopyFlow Gold for QuarkXPress
- DocBook
- Typo3 LocManager
- Icenix Infix (PDF)
- XLIFF source = target
- TXML Wordfast source = target

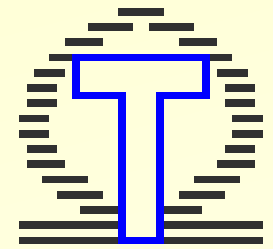
localization formats

- Android resources
- Java .properties
- Key-value files
- Mozilla DTD
- Windows resources (RC)
- WiX localisation
- ResX
- Flash XML export
- Camtasia for Windows
- Magento CE localisation
- PO (Portable Object File) (reading existing translations)
- SubRip subtitles (SRT)
- SVG images

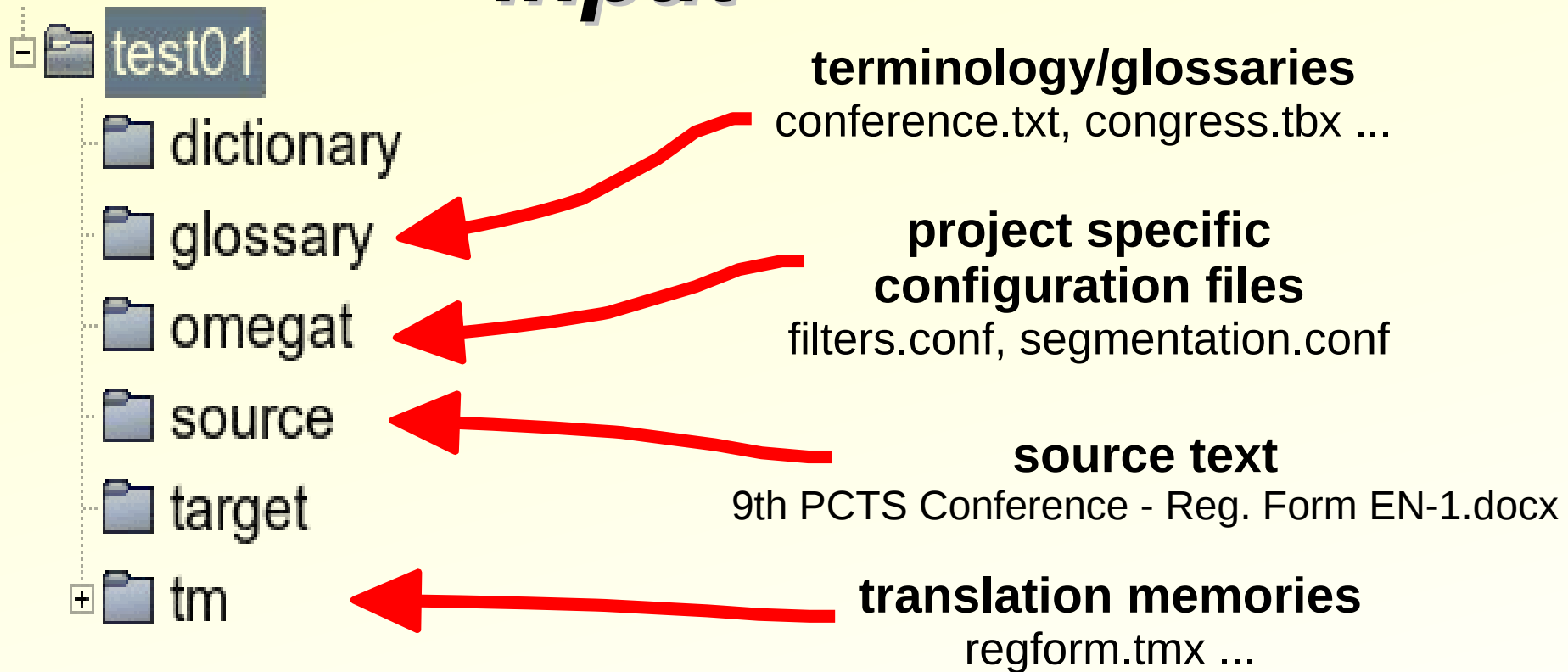
start OmegaT



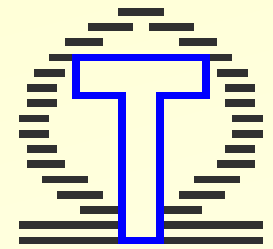
OmegaT: organisation



input



OmegaT: statistics

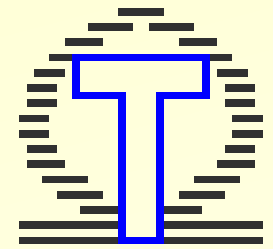


Project Statistics				
	Segments	Words	Characters (without spaces)	Characters (including spaces)
Total:	28	149	921	1031
Remaining:	16	68	424	559
Unique:	28	149	921	1031
Unique Remaining:	16	68	424	559

File Name	Total Segments	Remaining Segments
9th PCTS Conference - Reg. Form EN-1.docx	28	16

	Segments	Words	Characters (without spaces)	Characters (including spaces)
Repetitions:	0	0	0	0
Exact match:	12	81	497	559
95%-100%:	5	20	124	139
85%-94%:	1	7	27	33
75%-84%:	1	6	31	37
50%-74%:	4	15	102	112
No match:	5	20	140	151
Total:	28	149	921	1031

OmegaT: editor



The screenshot shows the OmegaT 3.1.4 editor window titled "test01". The main editor pane displays a document with several lines of text, some highlighted in yellow and green. The text includes contact information for Claudia Icobescu, Mihaela Visky, and Vasile Gherhe, followed by a paragraph about contacting organizers. Below this is a registration form with fields for "Vorname:", "First name:", and "Title:". The right-hand pane shows "Fuzzy Matches" with a list of matches, including "1. For further questions, do not hesitate to contact the organizers." and "Für zusätzliche Fragen steht Ihnen das Organisationsteam gerne zur Verfügung." Below the matches is a "Glossary" section with entries for "further questions = zusätzliche Fragen" and "organizers = Organisationsteam". The bottom status bar shows "Project autosaved on 5:43 PM" and page numbers "12/28 (12/28, 28)" and "65/77".

OmegaT-3.1.4 :: test01

Project Edit Go To View Tools Options Folders Help

Editor - 9th PCTS Conference - Reg. Form EN-1.docx

simona.simon@upt.ro

Claudia Icobescu <t0/> (für Vorträge in deutscher Sprache): <t1/>claudia.icobescu@upt.ro

Mihaela Visky <t0/> (für Vorträge in französischer Sprache): <t1/>mihaela.visky@upt.ro

Vasile Gherhe <t0/> (für Vorträge in rumänischer Sprache): <t1/>vasile.gherhes@upt.ro

For further questions, do not hesitate to contact the organizers.
Für zusätzliche Fragen steht Ihnen das Organisationsteam gerne zur Verfügung.
<segment 0010>

Informationen über den Teilnehmer

Vorname:

First name:

Title: <t0/> Mr/Ms/Mrs/Dr/Prof

Fuzzy Matches

1. For further questions, do not hesitate to contact the organizers.
Für zusätzliche Fragen steht Ihnen das Organisationsteam gerne zur Verfügung.
<100/100/100%
/home/c61302/temp/test01/tm/regform_ifaligner_en-de.tmx>

Glossary

further questions = zusätzliche Fragen

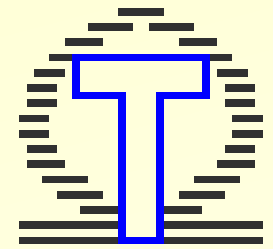
organizers = Organisationsteam

Dictionary Machine Translation Multiple Translations Notes Comments

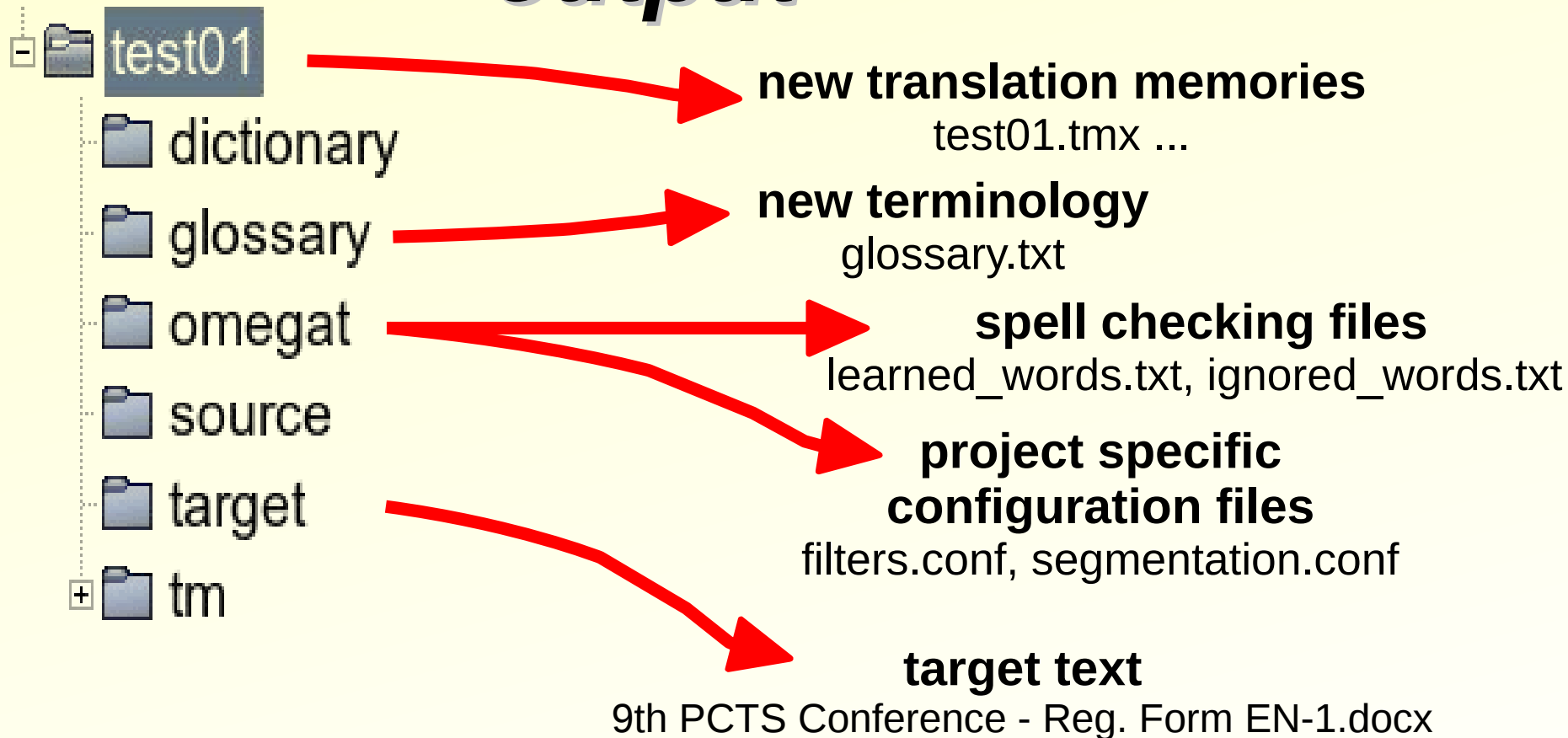
Project autosaved on 5:43 PM

12/28 (12/28, 28) 65/77

OmegaT: organisation



output



Part 1



Overview: part II

typical tasks of a translator

- ✓ **translate a website**
- ✓ create a TM on the basis of existing translations
- ✓ manage terminology and dictionaries
- ✓ extract terminology from texts
- ✓ use machine translation
- ✓ convert file formats
- ✓ manage bilingual files
- ✓ manage pdf files
- ✓ quality assessment
- ✓ use text corpora



▶ translate websites

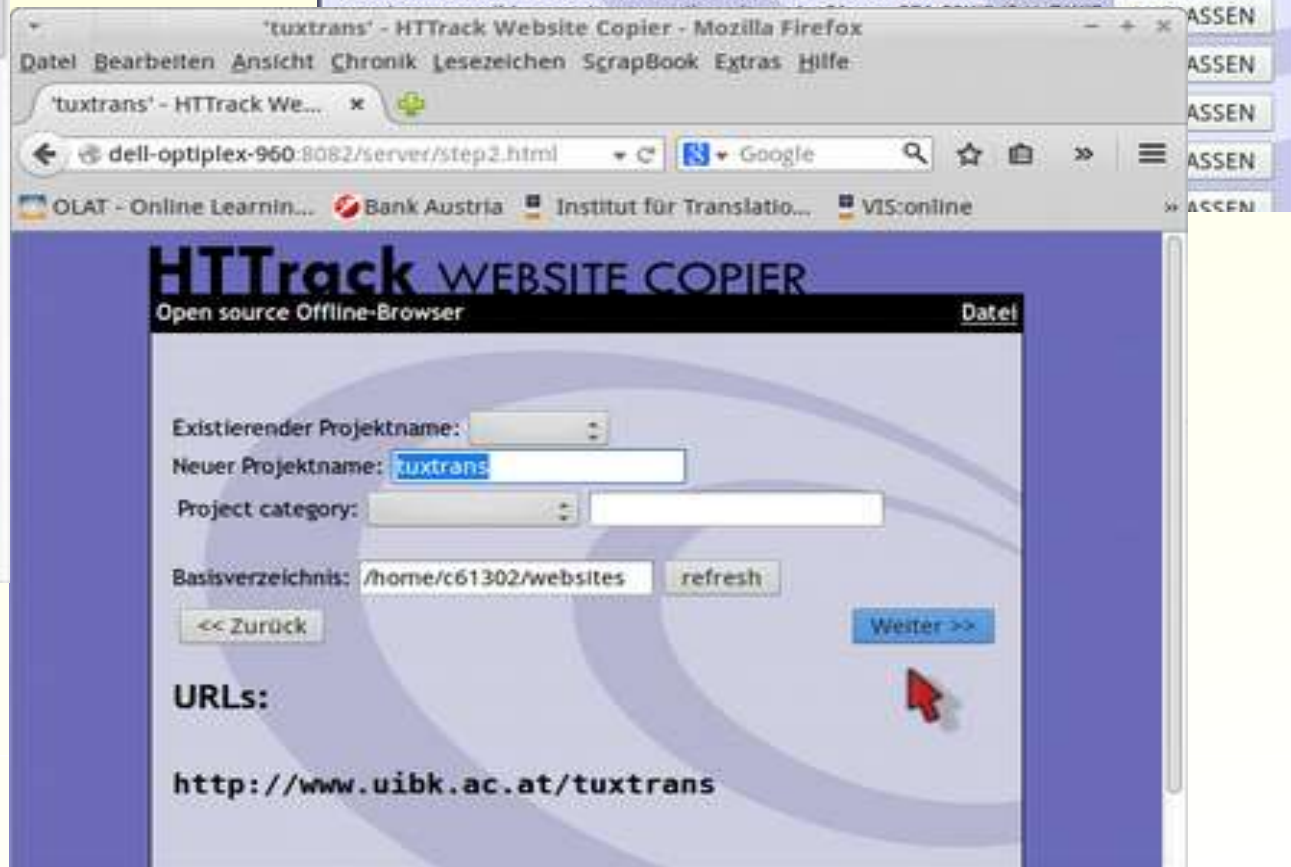
what you need:

- website with a few HTML documents
- local copy of website
- translation memory system
- existing translation memories
- existing terminology lists
- HTML editor



save website locally

Httrack website copier:



OmegaT: tuxtrans.org

OmegaT-3.1.5 - tuxtrans

Projekt Bearbeiten Gehe zu Ansicht Extras Optionen Folders Hilfe

Editor - tuxtrans/index.html

22de

22de

22de tuxtrans ist ein GNU/Linux Desktop-System für Übersetzer auf der Basis der Linux-Distribution Ubuntu

It is not just a full fledged operating system, though, it includes a collection of software applications which allow a translator to do his/her translation job most efficiently and in line with the latest standards.

Es ist nicht nur ein vollwertiges Betriebssystem, sondern beinhaltet auch eine Reihe von spezifischen Softwareanwendungen, die es Übersetzern erlauben, ihre Arbeit effizient sowie unter Anwendung der aktuellen Standards durchzuführen <Segment 2512>

tuxtrans: TuxTrans, Tuxtrans, PCLDSTrans, pclostrans, pclose-trans, PCLDS-Trans, linux for translators, computer aided translation, CAT, localization, localisation, machine aided translation, translation software, open source translation tools, translation tools, multilingual software, computer assisted translation, CAT, MAHT: machine aided human translation, linux, linux applications, linux live cd, linux operating system, linux os, linux software, linux system, localization, software localization, techtrans, translation, localization, translation memory, translator, translators, usb linux, live cd, live system

tuxtrans: Linux for Translators

css:tuxtrans.css

tux

Mehrfachübersetzungen Kommentare Notizen Wörterbuch Glossar

3/100 (2/1122, 3090) 216/236

1) But it is not just a full fledged operating system, it includes a collection of software applications which allow translation job most efficiently and in line with the latest standards. Es ist nicht nur ein vollwertiges Betriebssystem, sondern spezifischen Softwareanwendungen, die es Übersetzern sowie unter Anwendung der aktuellen Standards durchzuführen <100/94/93%> /home/c61302/Dropbox/zuschreiben/wien/beispiel/tuxtrans/

Maschinelle Übersetzung

Es ist nicht nur ein vollwertiges Betriebssystem, obwohl, enthält es eine Sammlung von Software-Anwendungen, die ein Übersetzer seine / ihre Übersetzungsarbeit möglichst effizient und im Einklang mit den neuesten Standards erlauben. <Google Translate v2>

400: Bad Request

09.2014 um Projektdateien

Dateiname	&Fil...	A...	N...
tuxtrans/software.html	HT...	199	128
tuxtrans/history.html	HT...	167	117
tuxtrans/credits.html	HT...	143	109
tuxtrans/apps/adat.html	HT...	70	69
tuxtrans/free.html	HT...	118	68
tuxtrans/install.html	HT...	126	67
tuxtrans/index.html	HT...	100	44
tuxtrans/apps/omegat.html	HT...	110	42
tuxtrans/get.html	HT...	87	37
tuxtrans/language.html	HT...	76	27
tuxtrans/start.html	HT...	81	27
tuxtrans/apps/anaphraseus...	HT...	78	26
tuxtrans/apps/pdftoolk.html	HT...	76	25
tuxtrans/apps/openproj.html	HT...	38	23
tuxtrans/apps/goldendict.html	HT...	70	20
tuxtrans/apps/kbabel.html	HT...	75	20
tuxtrans/apps/okapi.html	HT...	73	20
tuxtrans/apps/termbase.html	HT...	72	19
tuxtrans/apps/oit.html	HT...	78	17
tuxtrans/apps/projectlibre.h...	HT...	45	16
tuxtrans/apps/ubuntu.html	HT...	67	15
tuxtrans/apps/antconc.html	HT...	71	14
tuxtrans/apps/cmap.html	HT...	61	13
tuxtrans/apps/jubler.html	XHT...	33	13
tuxtrans/apps/textstat.html	HT...	68	13
tuxtrans/apps/craven.html	HT...	71	12
tuxtrans/apps/gaupol.html	HT...	65	11
tuxtrans/apps/poedit.html	HT...	73	11
tuxtrans/apps/toolkit.html	HT...	64	11
tuxtrans/apps/virtaal.html	HT...	70	11
tuxtrans/apps/alignassist.ht...	HT...	62	10
tuxtrans/apps/transolution....	XHT...	39	10
tuxtrans/apps/pdfsam.html	HT...	62	9
tuxtrans/apps/bitext2tmx.html	HT...	66	8
tuxtrans/apps/maxprograms...	HT...	58	8
tuxtrans/apps/subedit.html	HT...	65	7
tuxtrans/apps/Mc.html	HT...	59	7
tuxtrans/apps/gnotime.html	HT...	20	5
tuxtrans/apps/xournal.html	HT...	57	5
tuxtrans/apps/gpdftext.html	HT...	21	4
Gesamtzahl der Segmente		3.090	
Anzahl einmaliger Segmente		1.122	
Übersetzte einmalige Segm...			1

Move First

Move Up

Move Down

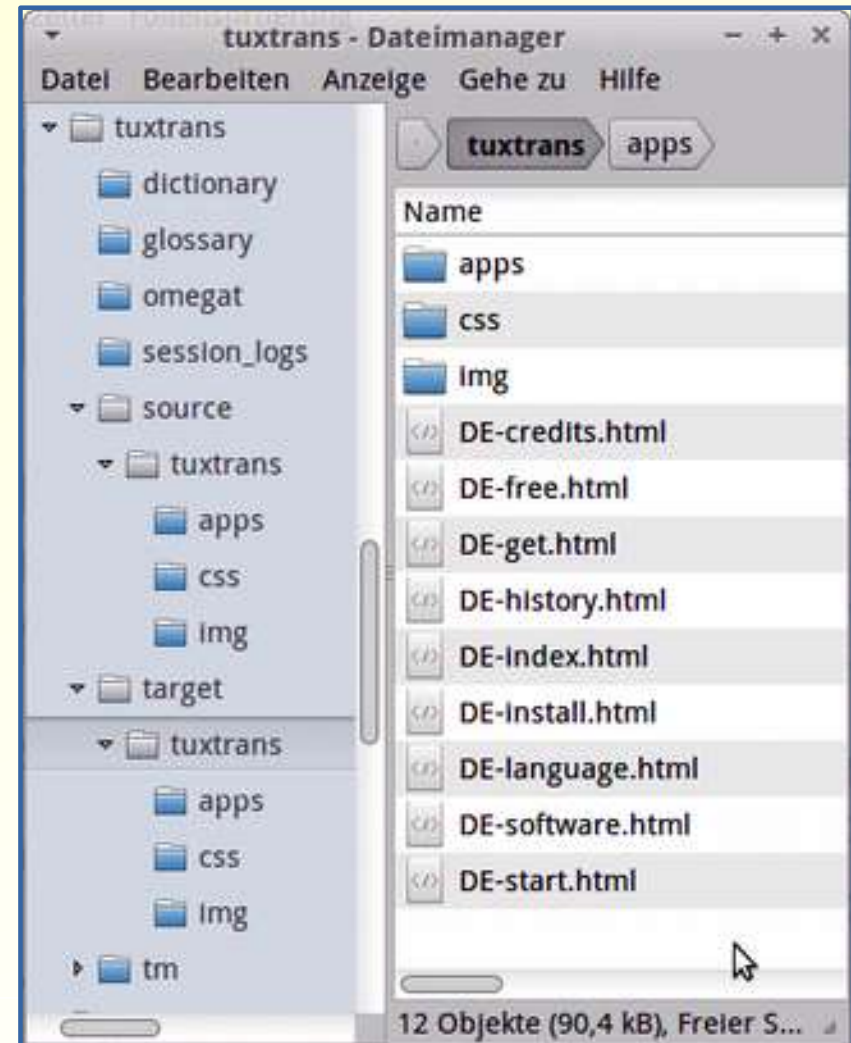
Move Last

Eine detailliertere Statistik befindet sich in der Datei: /home/c61302/Dropbox/zuschreiben/wien/beispiel/tuxtrans/omegat/project_stats.txt

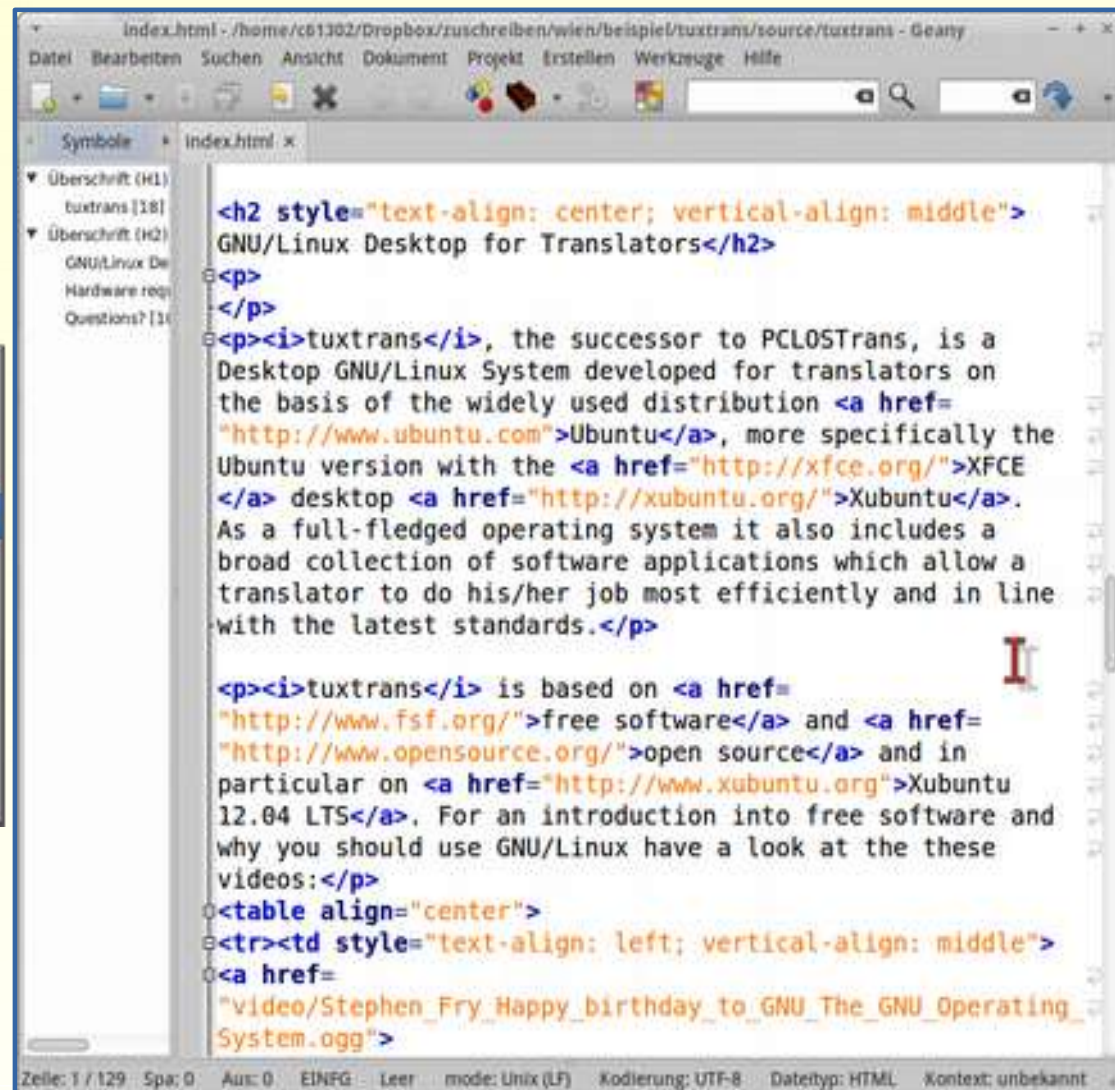
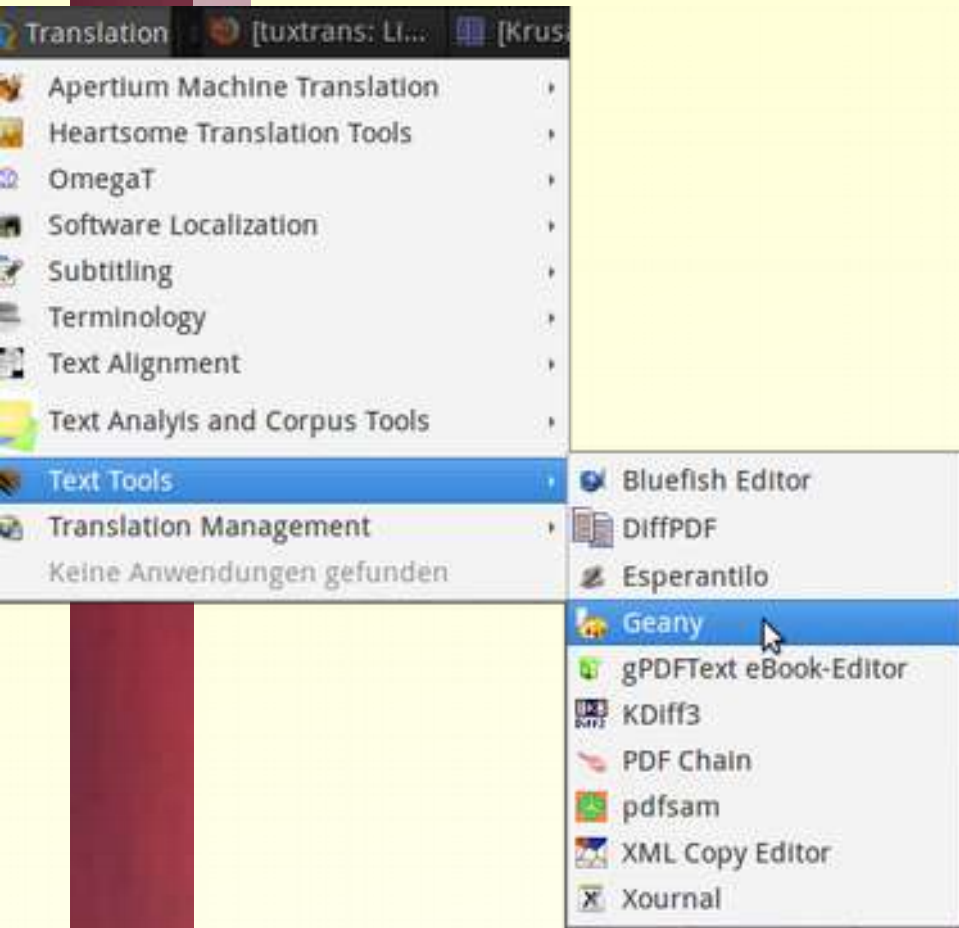
Dateien in den Quellordner kopieren... MediaWiki-Seite

OmegaT: tuxtrans.org

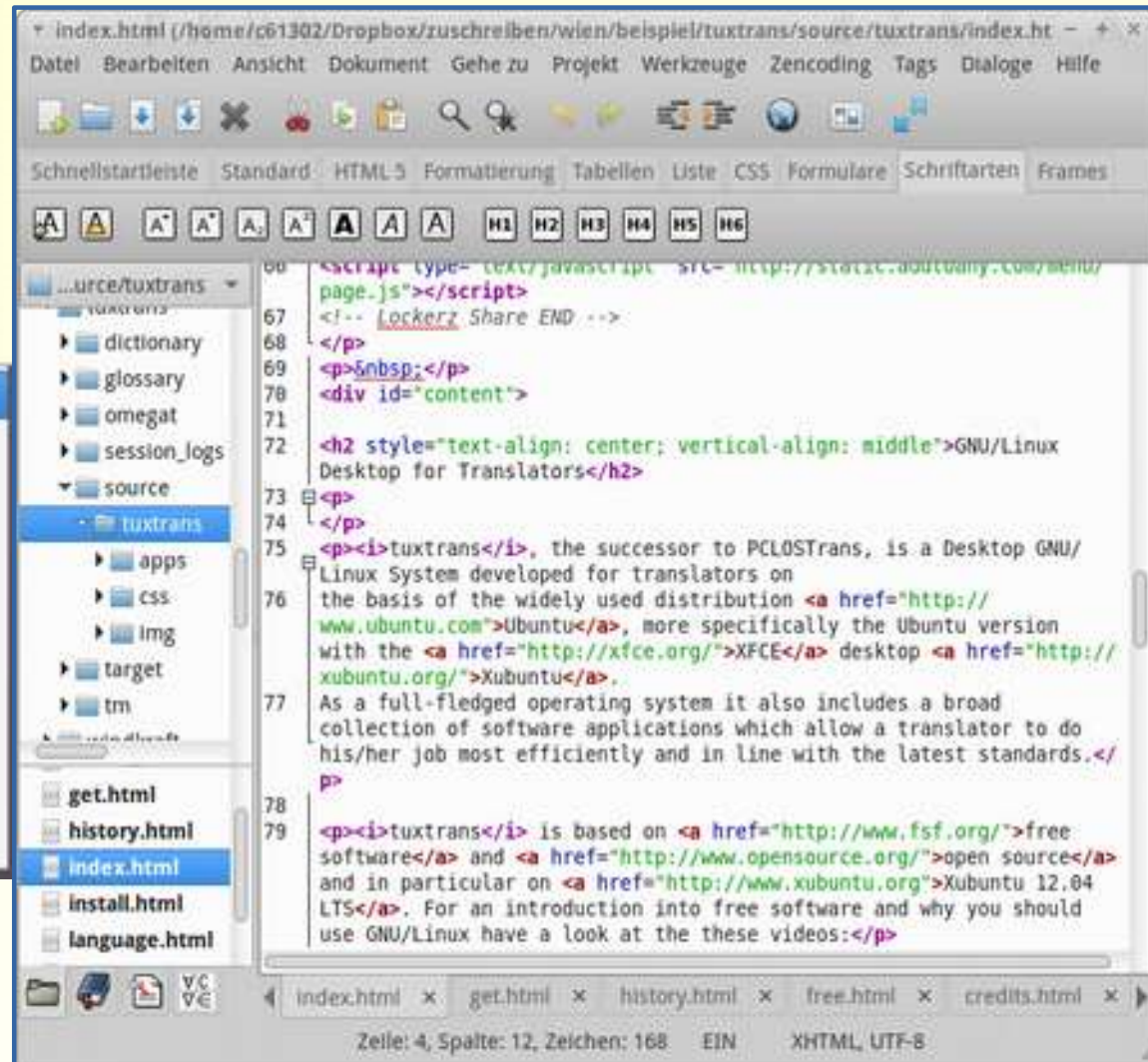
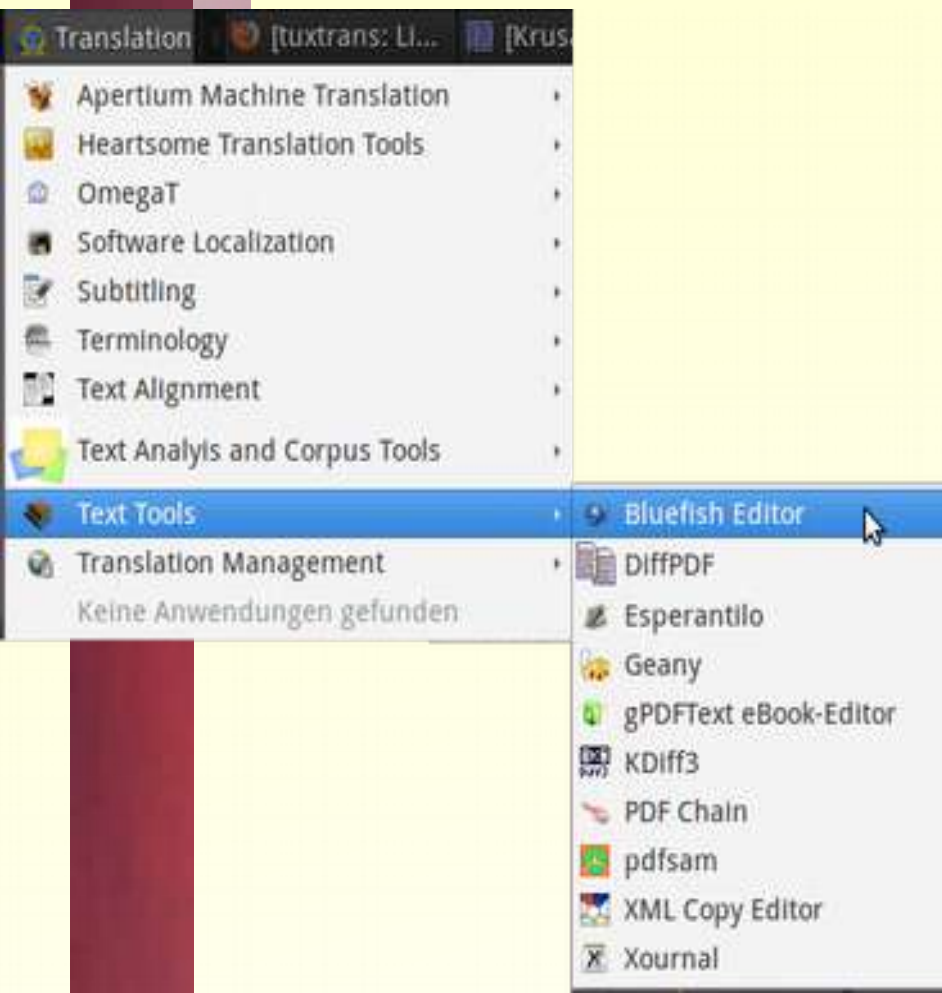
- source text = complete website directory structure with files
- target text = identical website directory structure with files



Edit HTML: Geany



Edit HTML: Bluefish



Overview: part II

typical tasks of a translator

- ✓ translate a website
- ✓ **create a TM on the basis of existing translations**
- ✓ manage terminology and dictionaries
- ✓ extract terminology from texts
- ✓ use machine translation
- ✓ convert file formats
- ✓ manage bilingual files
- ✓ manage pdf files
- ✓ quality assessment
- ✓ use text corpora



▶ Align source and target text

How can I create a translation memory from an existing translation?

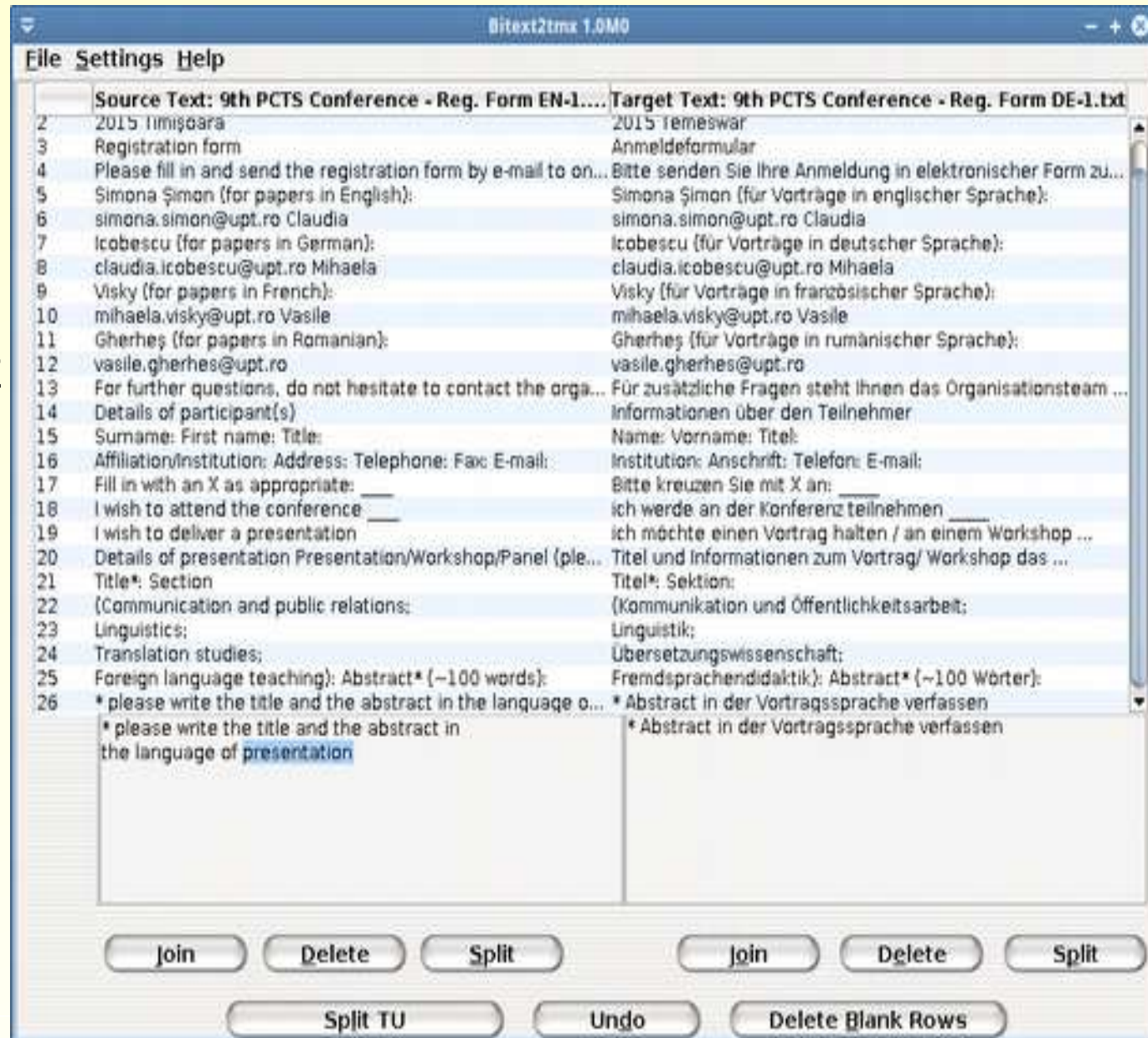
What you need:

- source text and target text
- an alignment tool
- time and patience



Alignment: BiText2TMX

- no development
- only text format
- TM as TMX



Alignment: LF-Aligner

- Formats:
txt (UTF-8!), rtf, doc,
docx, odt, pdf, html

9th PCTS Conference - Reg. Form EN-9th PCTS Conference - Reg. Form DE.xls - LibreOffice Calc

A	B
1 Instructions:	
2 1) Review and correct the pairings. See instructions on worksheet 2.	
3 2) Write your notes (to be added to each translation unit in the TMX) in column C if you wish.	
4 3) Save and close this file, and close any other open sheets.	
5 to the aligner window.	
6 International Conference On Professional Communication Translation Studies	Die 9. Internationale Konferenz im Bereich der Kommunikations- und Übersetzungswissenschaft
7 2015	26.-27.März 2015
8 in form	Temeswar
9 in and send the registration form by email to one of the following contact persons:	Anmeldeformular
10 Simon (for papers in English): simona.simon@upt.ro	Simona Simon (für Vorträge in englischer Sprache): simona.simon@upt.ro
11 Iacobescu (for papers in German): claudia.icobescu@upt.ro	Claudia Iacobescu (für Vorträge in deutscher Sprache): claudia.icobescu@upt.ro
12 Visky (for papers in French): mihaela.visky@upt.ro	Mihaela Visky (für Vorträge in französischer Sprache): mihaela.visky@upt.ro
13 Gherhes (for papers in Romanian): vasile.gherhes@upt.ro	Vasile Gherhes (für Vorträge in rumänischer Sprache): vasile.gherhes@upt.ro
14 For any questions, do not hesitate to contact the organizers.	Für zusätzliche Fragen steht Ihnen das Organisationsteam gerne zur Verfügung.
15 participant(s)	Informationen über den Teilnehmer
16 Name:	Vorname:
17 Title:	Name:
18 Institution:	Titel:
19 Address:	Institution:
20 Telephone: Fax: Email:	Anschrift:
21	Telefon: Fax: Email:
22 Fill in with an X as appropriate:	Bitte kreuzen Sie mit X an:
23 I wish to attend the conference	ich werde an der Konferenz teilnehmen
24	
25 I wish to deliver a presentation	ich möchte einen Vortrag halten / an einem Workshop teilnehmen / ein Podiumsgespräch halten
26	

```
Useless use of \E at ./scripts/LF_aligner_3.11_with_modules.pl line 51718.  
Useless use of \E at ./scripts/LF_aligner_3.11_with_modules.pl line 52054.
```

```
LF Aligner 3.11  
OS detected: Linux
```

Filetype?

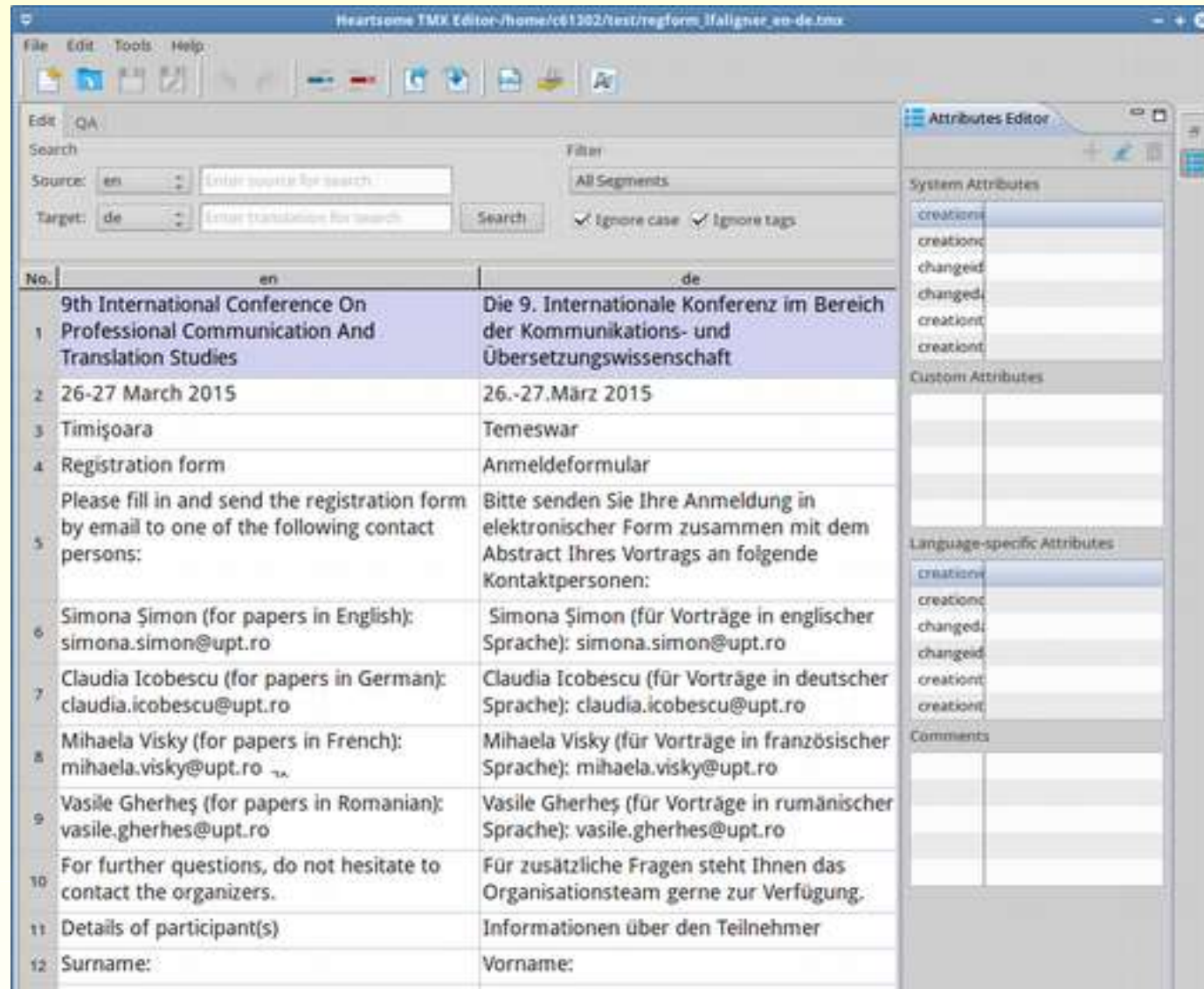
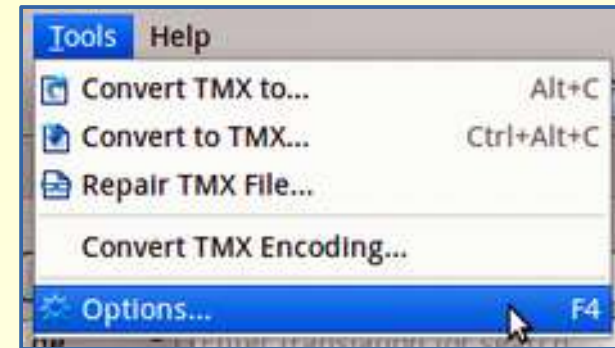
```
t - txt (UTF-8!), rtf, doc, docx or odt file (see the readme!)  
p - pdf, or pdf exported to txt (exporting works better, see readme!)  
h - HTML file saved to your computer  
w - webpage (you provide two URLs, the script does the rest)  
c - EU legislation by CELEX number (will be downloaded automatically)  
com - European Commission proposals (downloaded by year and number)  
epr - European Parliament reports (downloaded by year and number)
```

```
t/p/h/w/c/com/epr? (Default: t)
```


▶ TMX-Management

Heartsome TMX-Editor

- Edit, filter
- convert from TMX to ..., to TMX from ...
- QA
- ...



TMX-Management

Virtaal

- edit
- TMX QA

*regform_lfalligner_en-de.tmx - Virtaal

Datei Bearbeiten Anzeige Navigation Hilfe

Navigation: Alles

email to one of the following contact persons: elektronischer Form zusammen mit dem Abstract
Ihres Vortrags an folgende Kontaktpersonen:

Simona Şimon (for papers in English): simona.simon@upt.ro	Simona Şimon (für Vorträge in englischer Sprache): simona.simon@upt.ro
Claudia Icobescu (for papers in German): claudia.icobescu@upt.ro	Claudia Icobescu (für Vorträge in deutscher Sprache): claudia.icobescu@upt.ro
Mihaela Visky (for papers in French): mihaela.visky@upt.ro	Mihaela Visky (für Vorträge in französischer Sprache): mihaela.visky@upt.ro
Vasile Gherheş (for papers in Romanian): vasile.gherhes@upt.ro	Vasile Gherheş (für Vorträge in rumänischer Sprache): vasile.gherhes@upt.ro

For further questions, do not hesitate to contact the organizers. Für zusätzliche Fragen steht Ihnen das Organisationsteam gerne zur Verfügung.

Details of participant(s)		Einfache Mehrzahl(en), Klam...
Informationen über den Teilnehmer		

Surname:	Vorname:
First name:	Name: _
Title: _	Titel:
Affiliation/Institution:	Institution:
Address:	Anschrift:
Telephone: Fax: Email:	Telefon: Fax: Email: _
Fill in with an X as appropriate: _	Bitte kreuzen Sie mit X an:
I wish to attend the conference	ich werde an der Konferenz teilnehmen _
_ I wish to deliver a presentation	_ ich möchte einen Vortrag halten / an einem Workshop teilnehmen / ein Podiumsgespräch halten

Überprüft: Vorgabe Englisch → Deutsch

TMX-Management

Okapi Olifant

- edit
- QA
- alpha

Olifant (ALPHA) - defaultOlifantTMRepository

File Edit View Entries Translation Memory Help

Source: EN Target: DE Flagged entries only

regform_ifaligner_en-de

For further questions, do not hesitate to contact the organizers.

Für zusätzliche Fragen steht Ihnen das Organisationsteam gerne zur Verfügung.

Flag/SegKey	Text~EN	Text~DE
<input type="checkbox"/> 1	9th International Conference On Professional Communi	Die 9. Internationale Konferenz im Bereich der Kommuni
<input type="checkbox"/> 2	26-27 March 2015	26.-27.März 2015
<input type="checkbox"/> 3	Timișoara	Temeswar
<input type="checkbox"/> 4	Registration form	Anmeldeformular
<input type="checkbox"/> 5	Please fill in and send the registration form by email to o	Bitte senden Sie Ihre Anmeldung in elektronischer Form
<input type="checkbox"/> 6	Simona Şimon (for papers in English): simona.simon@up	Simona Şimon (für Vorträge in englischer Sprache): simo
<input type="checkbox"/> 7	Claudia Icobescu (for papers in German): claudia.icobesc	Claudia Icobescu (für Vorträge in deutscher Sprache): cla
<input type="checkbox"/> 8	Mihaela Visky (for papers in French): mihaela.visky@upt.	Mihaela Visky (für Vorträge in französischer Sprache): mil
<input type="checkbox"/> 9	Vasile Gherheş (for papers in Romanian): vasile.gherhes	Vasile Gherheş (für Vorträge in rumänischer Sprache): va
<input checked="" type="checkbox"/> 10	For further questions, do not hesitate to contact the org	Für zusätzliche Fragen steht Ihnen das Organisationstea
<input type="checkbox"/> 11	Details of participant(s)	Informationen über den Teilnehmer
<input type="checkbox"/> 12	Surname:	Vorname:
<input type="checkbox"/> 13	First name:	Name:
<input type="checkbox"/> 14	Title:	Titel:
<input type="checkbox"/> 15	Affiliation/Institution:	Institution:
<input type="checkbox"/> 16	Address:	Anschrift:
<input type="checkbox"/> 17	Telephone: Fax: Email:	Telefon: Fax: Email:
<input type="checkbox"/> 18	Fill in with an X as appropriate:	Bitte kreuzen Sie mit X an:

10 / 29 page 1 of 1

Overview: part II

typical tasks of a translator

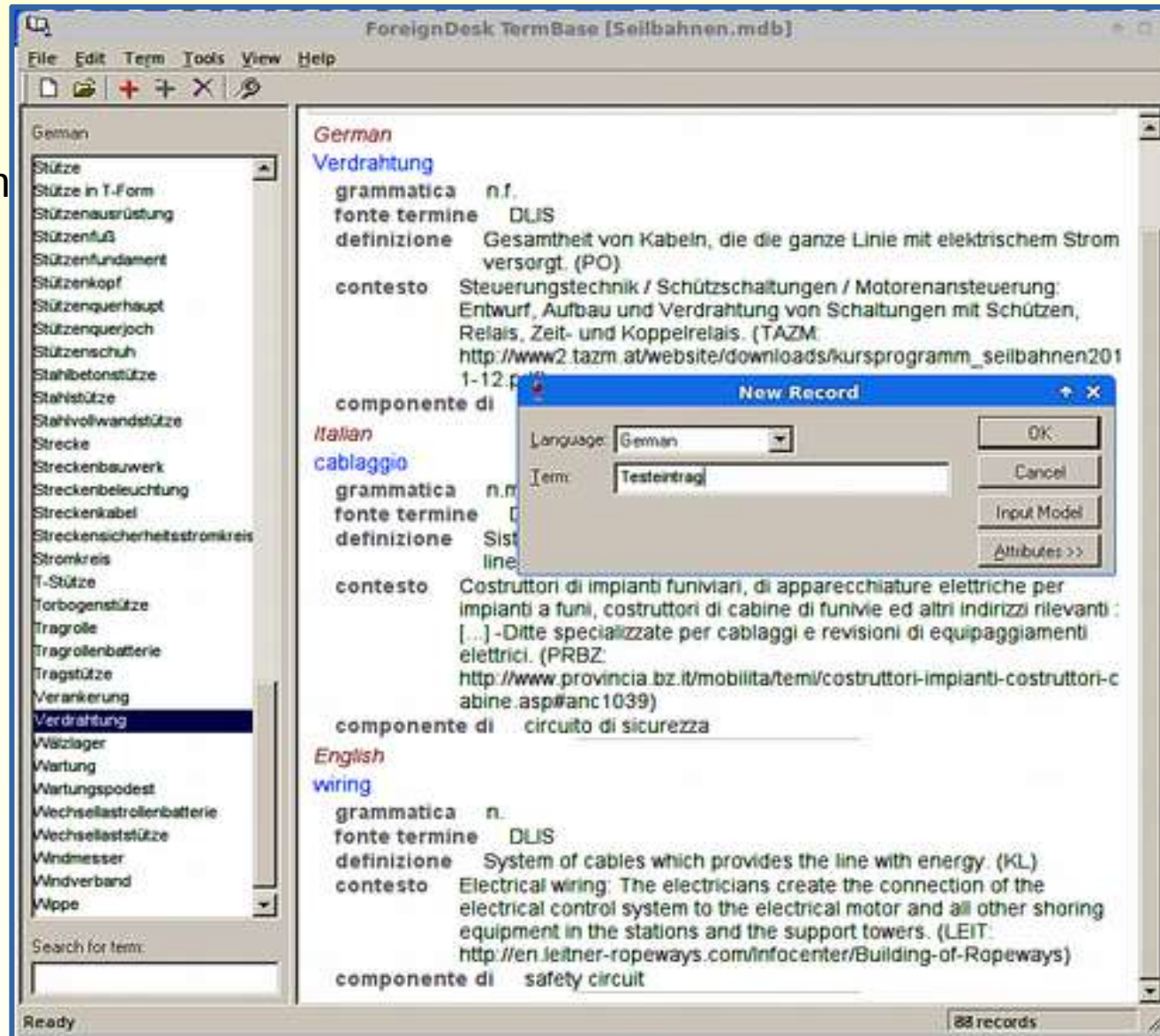
- ✓ translate a website
- ✓ create a TM on the basis of existing translations
- ✓ **manage terminology and dictionaries**
- ✓ extract terminology from texts
- ✓ use machine translation
- ✓ convert file formats
- ✓ manage bilingual files
- ✓ manage pdf files
- ✓ quality assessment
- ✓ use text corpora



► manage your terminology

ForeignDesk Termbase

- concept oriented
- import/export
html, csv, Multiterm
- OmegaT integration with
csv export +
edit



▶ search for terminology

GoldenDict

- several dictionary formats
- same formats as dictionaries in OmegaT

The screenshot shows the GoldenDict application window titled "wind power - GoldenDict". The search bar contains "wind power". The main content area displays the following information:

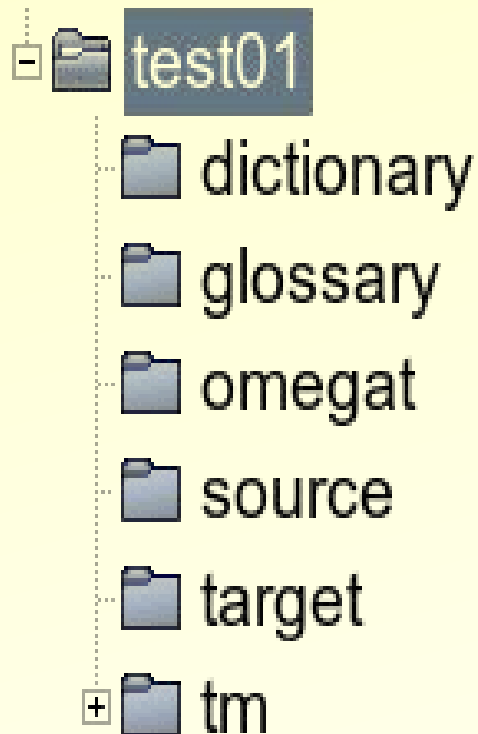
- Wikipedia**: English [edit]
- Noun [edit]**: **wind power** (*uncountable*)
 1. Power harnessed from the wind.
- Translations [edit]**: **power harnessed from the wind**
 - Catalan: **energia eòlica**
 - Chinese: Mandarin: 风能 (zh), 风能 (fēngnéng), 风力 (zh), 风力 (fēnglì)
 - Danish: **vindenergi** (da)
 - Dutch: **windenergie** (nl) f
 - Esperanto: **ventoenergio**
 - Estonian: **tuuleenergia**
 - Finnish: **tuulivoima** (fi), **tuulisähkö** (fi)
 - French: **énergie éolienne** f
 - German: **Windenergie** (de) f
 - Icelandic: **vindorka**
 - Italian: **energia eolica** f
 - Japanese: 風力 (ja) (ふうりょく, fūryoku)
 - Russian: **энергия ветра** f (energija vétra)
 - Spanish: **energía eólica** (es)
 - Swedish: **vindkraft** (sv)

On the right side, a sidebar titled "In Wörterbüchern gefunden:" lists search results from various sources:

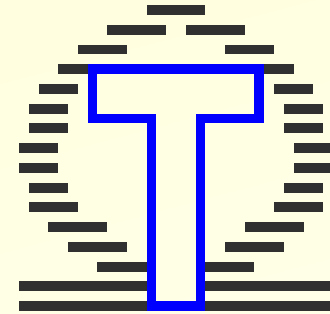
- English Wikipedia
- English Wiktionary
- German Wiktionary

At the bottom right, a "Verlauf:" (History) section shows a list of search terms: "wind power", "wind energy", "Windenergie", "windkraft", "energy", "Gefäßkrampfs", and "Gefäßbündeln".

integrate dictionaries in OmegaT



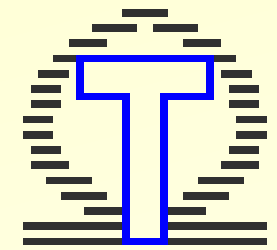
dictionaries



dictionaries vs glossaries:

- no interaction possible
just reading of entries
no editing nor adding
- all dictionaries in stardict format
- available on the web

dictionaries in OmegaT



monolingual StarDict dictionary
integrated in OmegaT:
„Oxford Advanced Learner dictionary“

The screenshot shows the OmegaT-3.1.4 interface. The main editor window displays a document titled "9th PCTS Conference - Reg. Form EN-1.docx" with the following text:

Bitte kreuzen Sie mit X an:

ich werde an der Konferenz teilnehmen

ich möchte einen Vortrag halten / an einem Workshop teilnehmen / ein Podiumsgespräch halten

Details of presentation

Presentation/Workshop/Panel (please indicate):

Title <t0/>*<t1/>:

Section (Communication and public relations; Linguistics; <t0/> Translation studies; Foreign language teaching <t1/>): <segment 0026>

Abstract* (~100 words):

* please write the title and the abstract in the language of presentation

The "Fuzzy Matches" panel shows a match for "1. Communication and Kommunikation und" with a score of <36/36/24% and a path to a dictionary file: /home/c61302/temp-1-de.tmx>

The "Dictionary" panel shows the definition for "Linguistics":

Linguistics - Study of the nature and structure of language. Linguists use a synchronic (describing a language as it exists at a given time) or a diachronic (tracing a language's development through its history) approach to language study. Greek philosophers in the 5th cent. BC who debated the origins of human language were the first in the West to be concerned with linguistic theory. The first complete Greek grammar, written by Dionysus Thrax in the 1st cent. BC, was a model for Roman grammarians, whose work led to the medieval and Renaissance vernacular grammars. With the rise of historical linguistics in the 19th cent., linguistics became a science. In the late 19th and early 20th cent., F. Saussure established the structuralist school of linguistics, which analyzed actual speech to learn about the underlying structure of language. In the 1950s, N. Chomsky challenged the structuralist approach, arguing that linguistics should study native speakers' unconscious knowledge of their own language (competence), not their actual production of language (performance), and developed generative grammar.

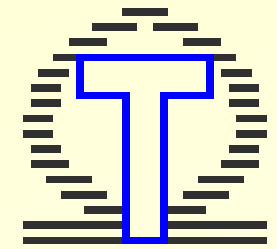
The "Glossary" panel shows the following entries:

linguistics = Sprachwissenschaft

Section = Sektion

The bottom status bar shows "Machine Translation Multiple Translations Notes Comments" and "Project autosaved on 10:30 AM". The bottom right corner displays "18/28 (18/28, 28) 116/0".

dictionaries in OmegaT

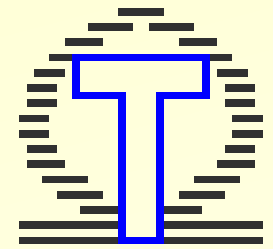


bilingual StarDict dictionaries
in OmegaT: „en-de“

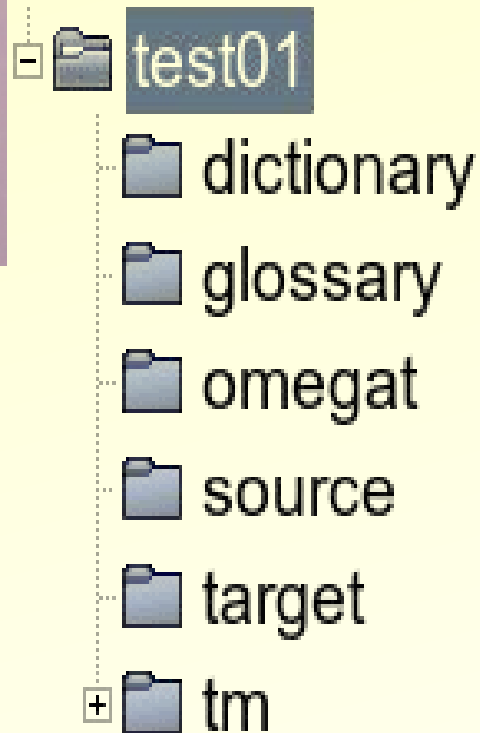
The screenshot shows the OmegaT 3.1.4 interface with the following components:

- Editor:** Displays a document titled "9th PCTS Conference - Reg. Form EN-1.docx". The text includes names and contact information in German and Romanian, such as "Claudia Iacobescu" and "Mihaela Visky".
- Fuzzy Matches:** Shows a list of matches, including "1. For further quest the organizers." and "Für zusätzliche Fra Organisationsteam".
- Dictionary:** Displays a list of dictionary entries for the word "contact" and "further", including parts of speech and German translations.
- Glossary:** Shows a list of terms and their translations, such as "further questions = zusätzliche Fragen" and "organizers = Organisationsteam".
- Machine Translation:** Shows the translated text, including "Für zusätzliche Fragen steht Ihnen das Organisationsteam gerne zur Verfügung."

glossaries in OmegaT



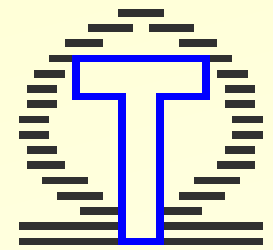
create, edit and use glossary lists in OmegaT



glossaries

- simple 3-column format
source term (Tab) target term (Tab) note
- always UTF-8 coded
- file names: *.txt, *.tab, *.utf8, *.tbx
- terminology recognition and auto-completer in OmegaT:
<https://www.youtube.com/watch?v=LszlaP22QhQ>

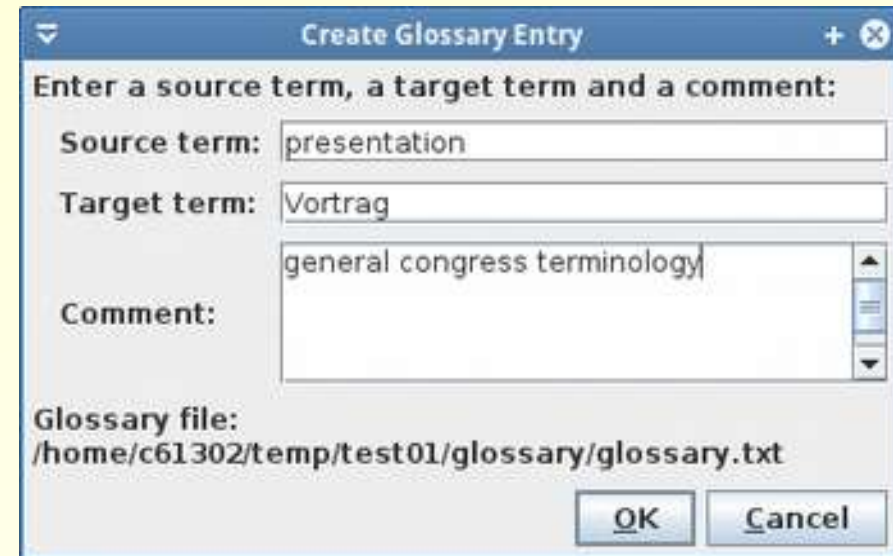
glossaries in OmegaT



create, edit and use glossary lists in OmegaT

- add terminology:
[CTRL Shift G] or
Edit / Create glossary entry

- saved in
Project / Properties:



Create Glossary Entry

Enter a source term, a target term and a comment:

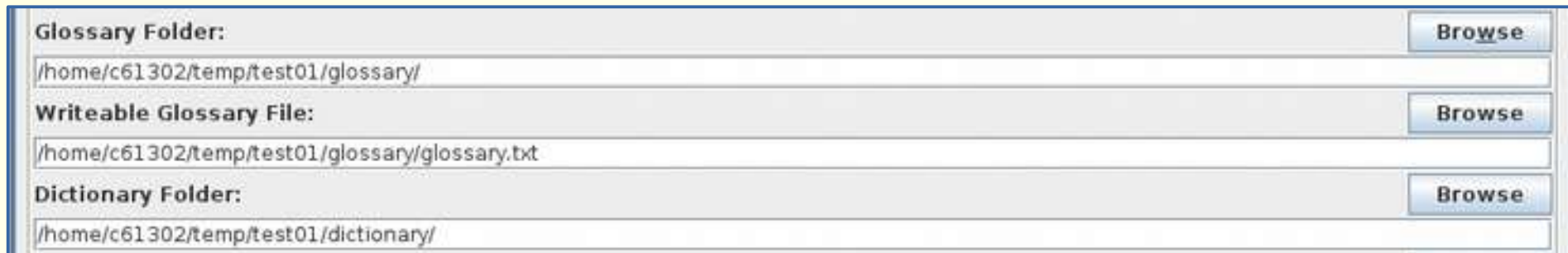
Source term: presentation

Target term: Vortrag

Comment: general congress terminology

Glossary file:
/home/c61302/temp/test01/glossary/glossary.txt

OK Cancel



Glossary Folder: /home/c61302/temp/test01/glossary/ Browse

Writeable Glossary File: /home/c61302/temp/test01/glossary/glossary.txt Browse

Dictionary Folder: /home/c61302/temp/test01/dictionary/ Browse

Overview: part II

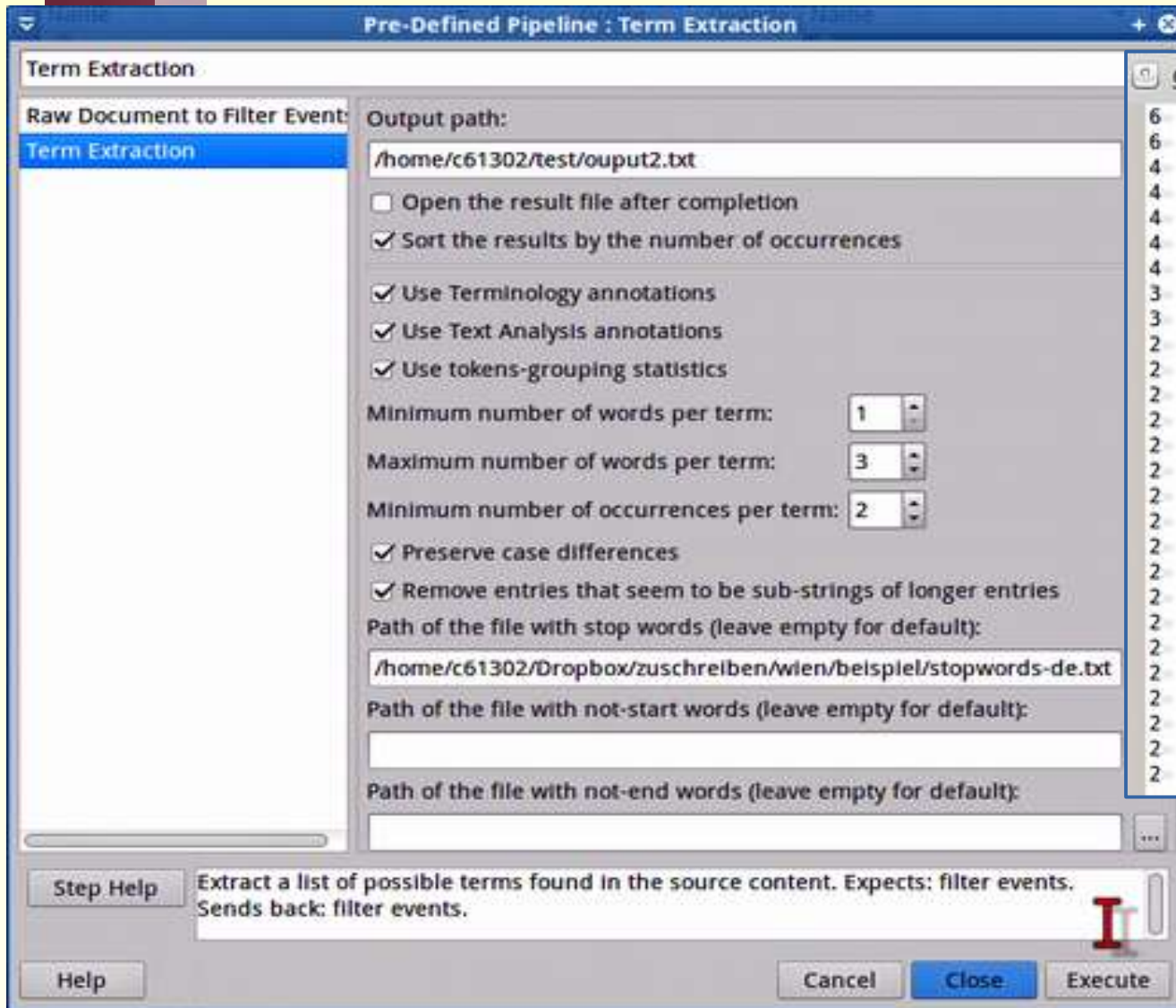
typical tasks of a translator

- ✓ translate a website
- ✓ create a TM on the basis of existing translations
- ✓ manage terminology and dictionaries
- ✓ **extract terminology from texts**
- ✓ use machine translation
- ✓ convert file formats
- ✓ manage bilingual files
- ✓ manage pdf files
- ✓ quality assessment
- ✓ use text corpora



▶ term extraction

- simple monolingual term extraction with Rainbow

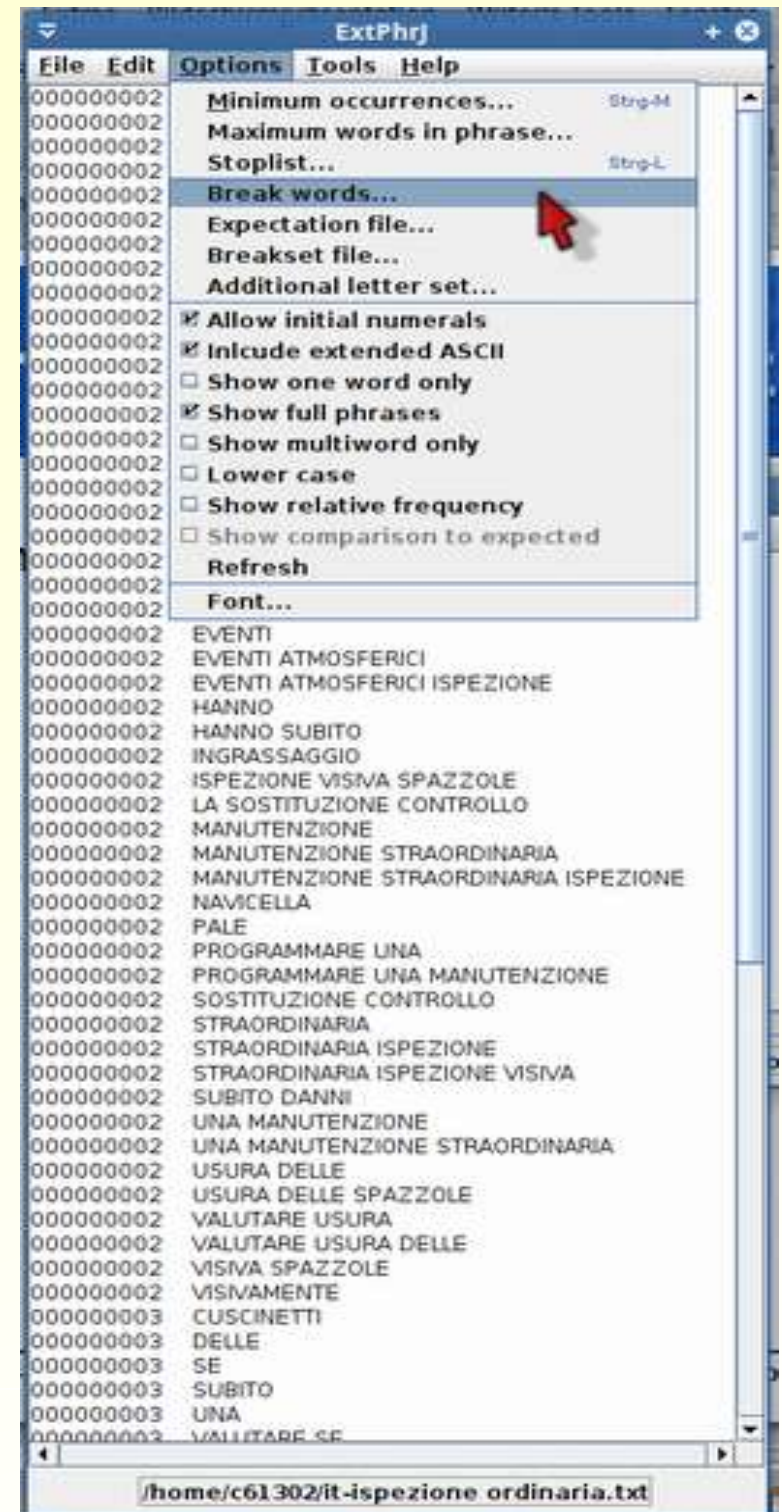


```
ouput.txt (Viewing) X
6 caso
6 programmare
4 Ispezione vi
4 sostituire p
4 sostituzione
4 usura sostit
4 visiva
3 Valutare
3 usura
2 CONTROLLO
2 ISPEZIONE VIS
2 ISPEZIONE VIS
2 SPAZZOLE
2 VISIVA
2 VISIVA SPAZZO
2 Valutare usu
2 atmosferici
2 causa
2 cuscinetti
2 danneggiament
2 danni causa
2 eventi atmos
2 manutenzione
2 spazzole
2 straordinari
2 subito danni
2 visivamente
```

```
ouput2.txt (Viewing) X
7 Sichtprüfung
6 veranlassen
4 Austausch veranlassen
4 Bei Verschleiß ersetzen
4 Kohlebürsten
4 Prüfen
4 Verschleiß ersetzen
4 ersetzen
3 DER
3 Einfetten
3 SICHTPRÜFUNG DER
2 Elektrokabel
2 Falle
2 Im Falle
2 Instandhaltung veranlassen
2 Kugellager
2 Schwingungsdämpfer
2 Schäden außerordentliche
2 Instandhaltung
2 Verschleißspuren aufweisen
2 Verschleißzustand
2 Witterungsereignisse
2 beschädigt
2 aufweisen
2 außerordentliche
2 Instandhaltung veranlassen
2 beschädigt
2 entsprechenden
2 prüfen
2 Überprüfen
```


▶ term extraction

- simple monolingual term extraction with Phrase extractor



Overview: part II

typical tasks of a translator

- ✓ translate a website
- ✓ create a TM on the basis of existing translations
- ✓ manage terminology and dictionaries
- ✓ extract terminology from texts
- ✓ **use machine translation**
- ✓ convert file formats
- ✓ manage bilingual files
- ✓ manage pdf files
- ✓ quality assurance
- ✓ use text corpora



▶ Machine Translation

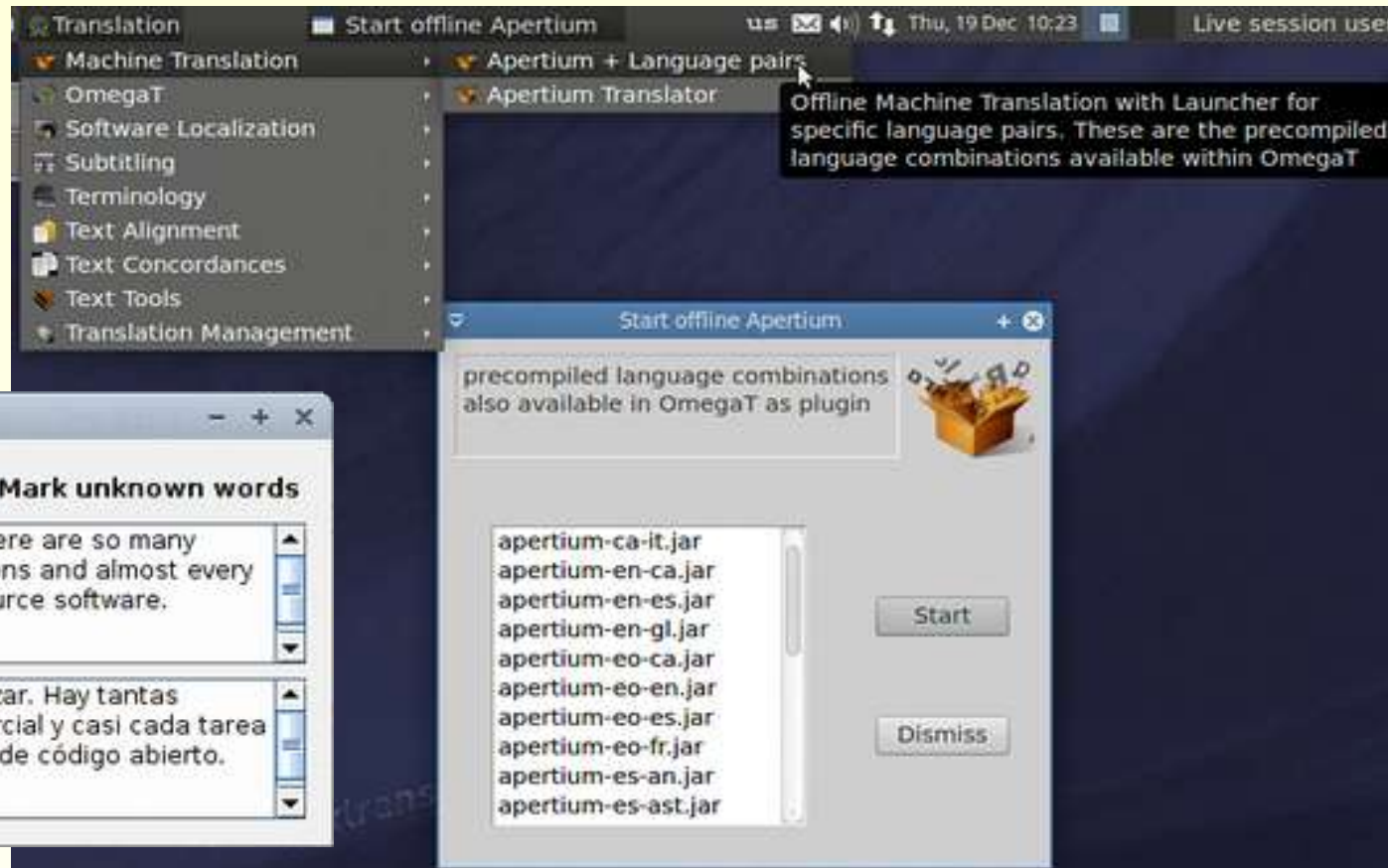
what you need:

- an Open Source MT system
- a free on-line MT system
- an interface to your TM system



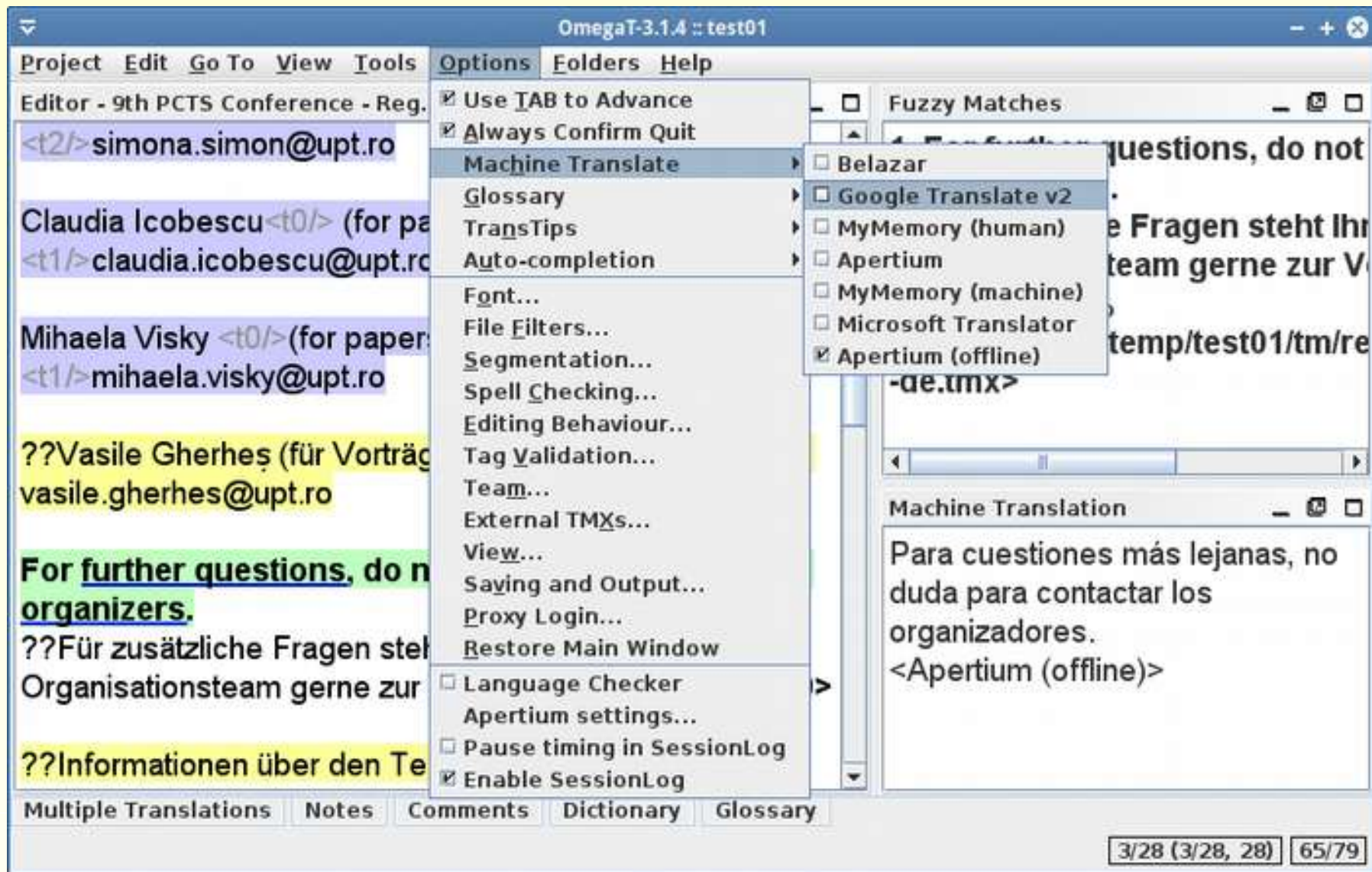
Apertium: OS MT system

- Apertium is a rule based Open Source MT system
- mostly regional language pairs from Spain with English

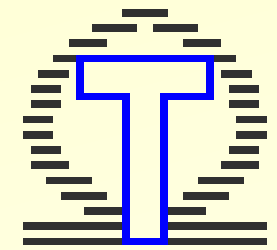


Apertium: OS MT system

- Apertium offline can be integrated into OmegaT



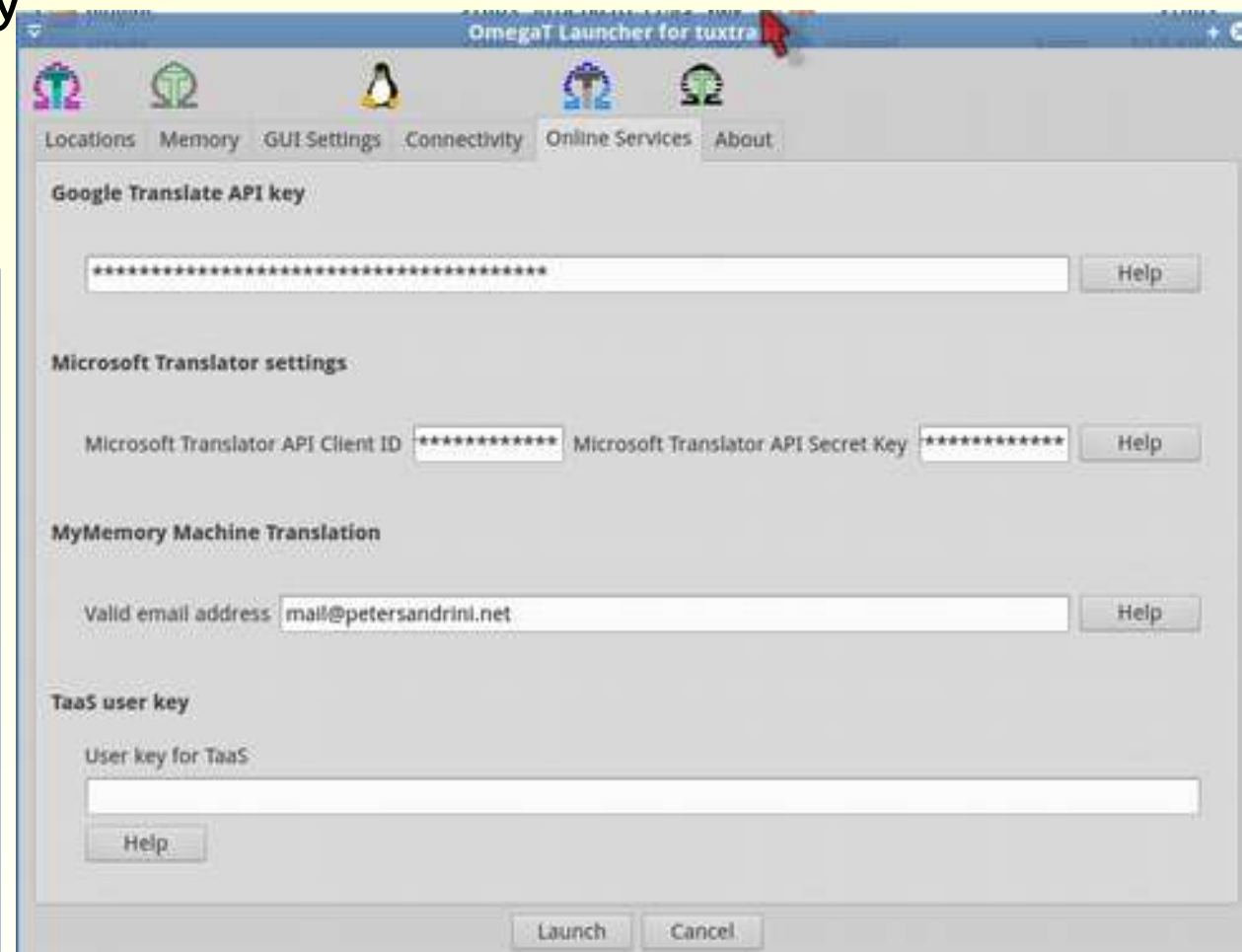
on-line MT systems



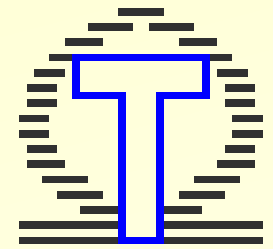
- you need to register to use online Mt within OmegaT:
 - Mymemory (only E-Mail)
 - Microsoft Translator API ID + Secret key
 - Google Translate API key

OmegaT.I4J.ini

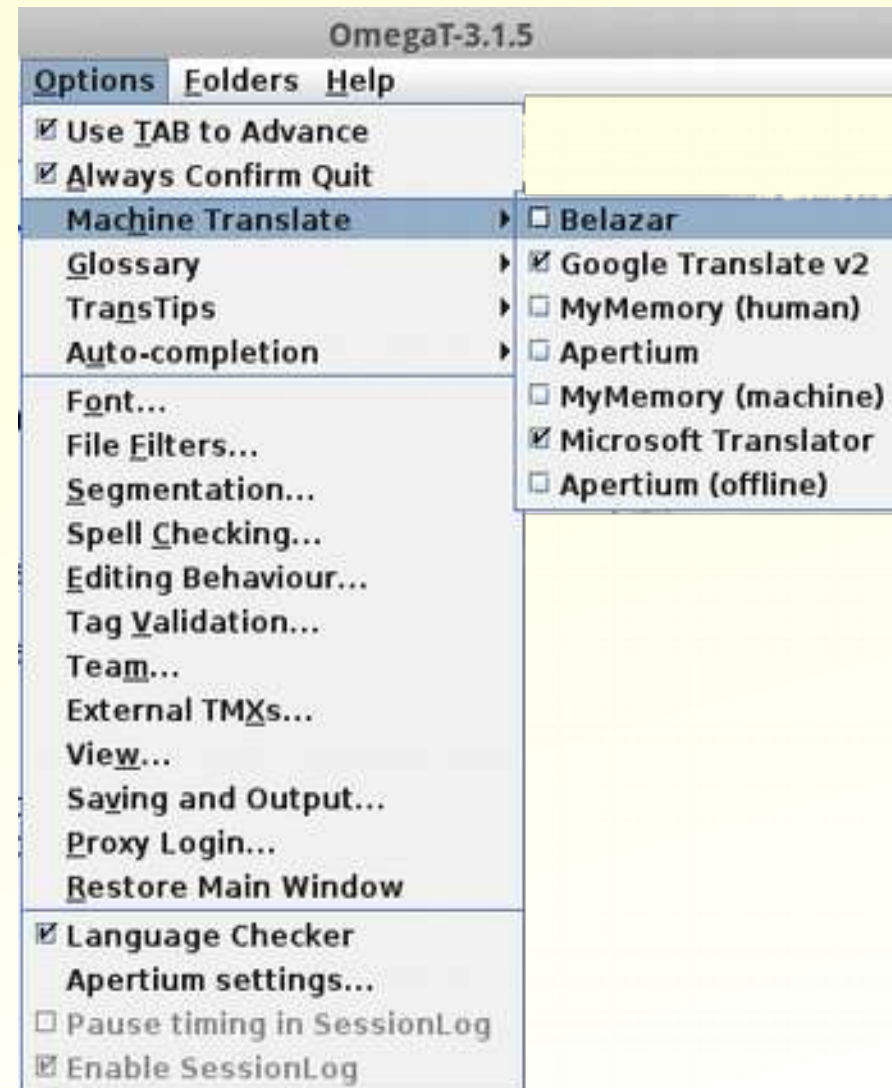
```
# OmegaT.exe runtime configuration
# To use a parameter, remove the '#' before the '-'
# Memory
-Xmx512M
# Language
-Duser.language=en
# Country
#-Duser.country=AT
# Settings to access the Internet behind a proxy
#-Dhttp.proxyHost=192.168.1.1
#-Dhttp.proxyPort=3128
# Google Translate v2 API key
#-Dgoogle.api.key=xxxxx
# Microsoft Translator credentials
#-Dmicrosoft.api.client_id=xxxxx
#-Dmicrosoft.api.client_secret=xxxxx
# MyMemory email
#-Dmymemory.api.email=xxxxx@xxxxx.xx
# TaaS user key
#-Dtaas.user.key=xxxxx
```



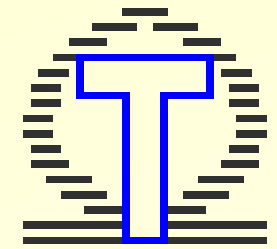
on-line MT systems



- available for free on the web
 - Microsoft Translator
 - Google Translate
- within OmegaT
 - Apertium online + offline free
 - Mymemory free
 - Microsoft Translator free, but you need to register
 - Google Translate registration + costs



on-line MT systems



- interface in OmegaT

The screenshot shows the OmegaT-3.1.5 interface with the following components:

- Window Title:** OmegaT-3.1.5 :: tuxtrans
- Menu Bar:** Projekt, Bearbeiten, Gehe zu, Ansicht, Extras, Optionen, Folders, Hilfe
- Editor - tuxtrans/index.html:**
 - Original text (German): Sie können andere Desktops oder Window-Manager leicht aus den Ubuntu Software-Repositories installieren.
 - Segment: **<i1>tuxtrans</i1> uses Xubuntu 12.04, the <a2>long term support</a2> (LTS) version, as its basis which guarantees 5 years support and updates as well as a stable two year release cycle.**
 - Target text (English): **<i1>tuxtrans</i1> Xubuntu 12.04 verwendet, die <a2>langfristige Unterstützung</a2> (LTS)-Version, als seine Basis, der 5 Jahre Support und Updates sowie einen stabilen Release-Zyklus 2 Jahre garantiert.<Segment 0896>**
 - Link: [software.html](#)
 - Footnote: **<i0>tuxtrans</i0> comes with a lot of applications suited to the everyday tasks of the translator or everybody dealing with multilingual texts.**
- Maschinelle Übersetzung:**
 - Machine translation output: **<i1>tuxtrans</i1> Xubuntu 12.04 verwendet, die <a2>langfristige Unterstützung</a2> (LTS)-Version, als seine Basis, der 5 Jahre Support und Updates sowie einen stabilen Release-Zyklus 2 Jahre garantiert. <Google Translate v2>**
 - Another machine translation output: **<i1>tuxtrans</i1> Xubuntu 12.04 verwendet, die <a2>langfristige Unterstützung</a2> (LTS)-Version, als seine Basis, der 5 Jahre Support und Updates sowie einen stabilen Release-Zyklus 2 Jahre garantiert. <MyMemory (machine)>**
 - Error message: **400: Bad Request <Microsoft Translator>**
- Footer:** Mehrfachübersetzungen, Kommentare, Notizen, Wörterbuch, Glossar, Unschärfe Treffer, 7/100 (6/1122, 3090), 185/202

Overview: part II

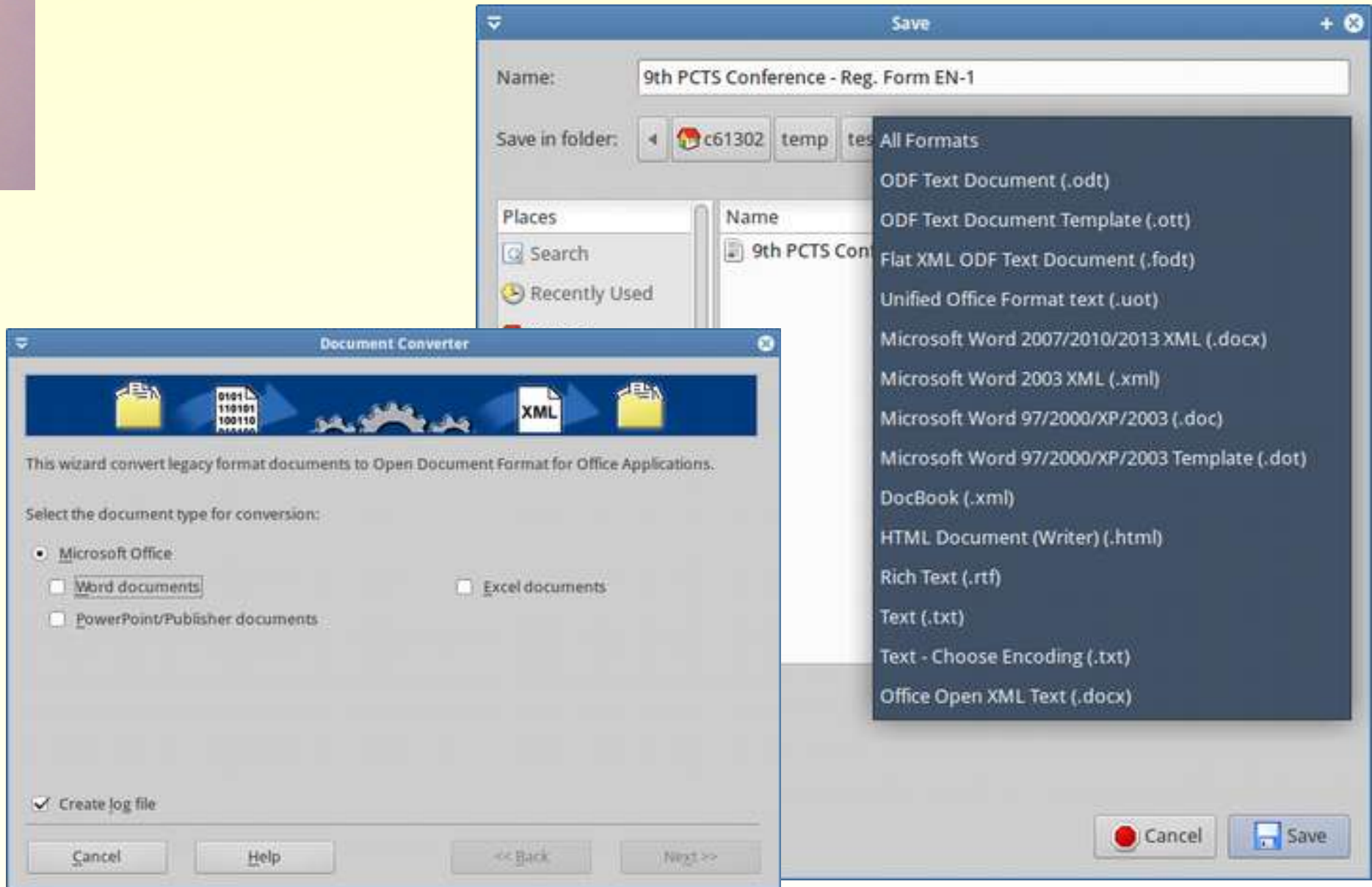
typical tasks of a translator

- ✓ translate a website
- ✓ create a TM on the basis of existing translations
- ✓ manage terminology and dictionaries
- ✓ extract terminology from texts
- ✓ use machine translation
- ✓ **convert file formats**
- ✓ manage bilingual files
- ✓ manage pdf files
- ✓ quality assurance
- ✓ use text corpora



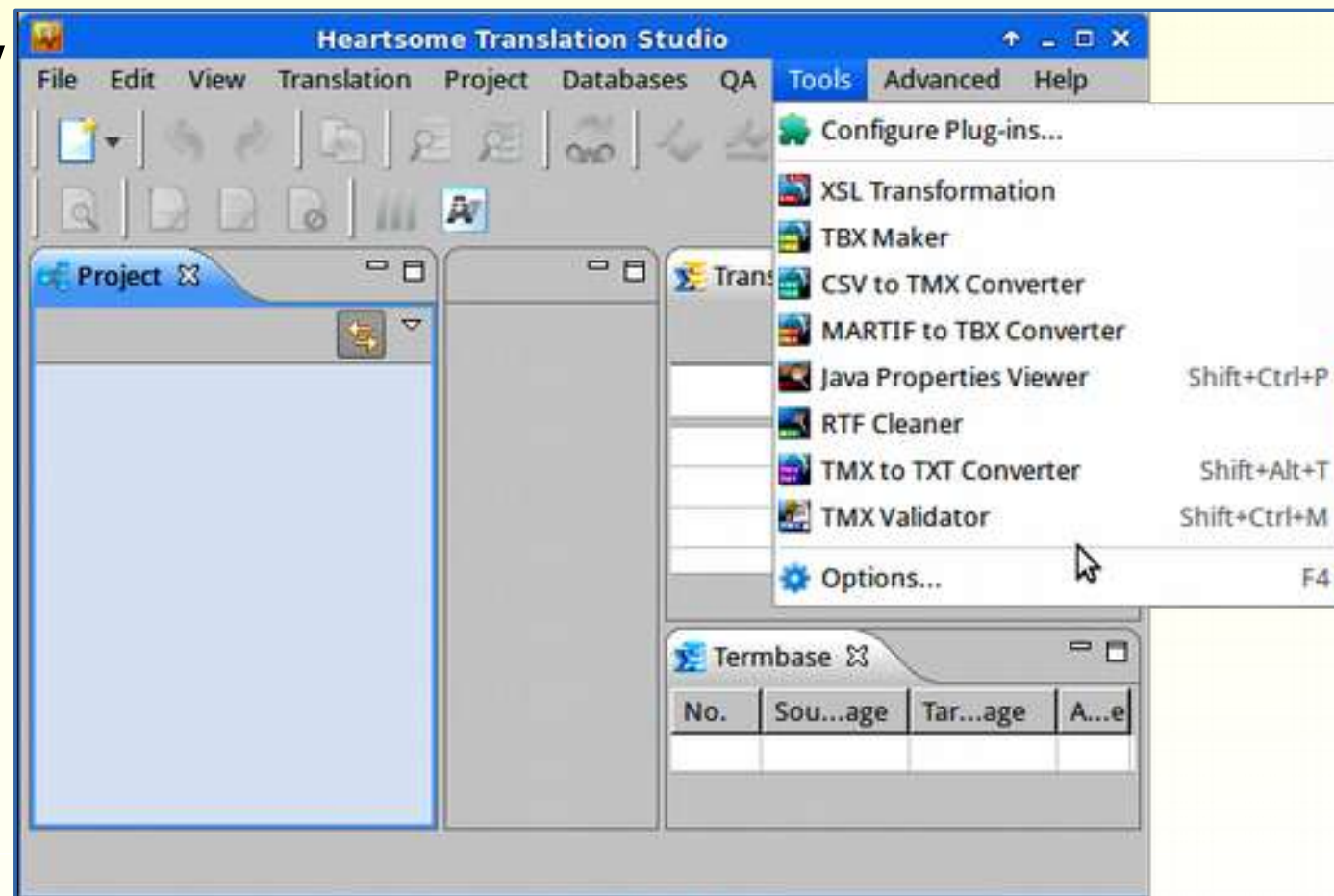
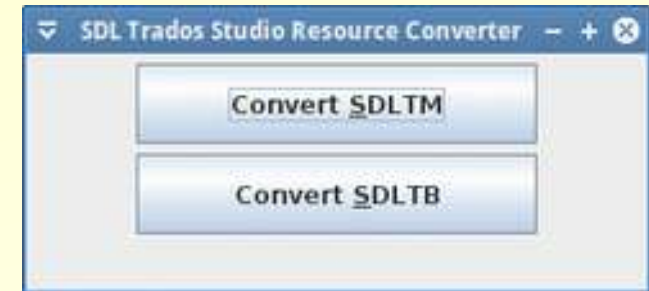
convert file formats

- text formats from doc to docx, odt ... with LibreOffice



convert file formats

- CSV, TAB, TXT glossary lists to TBX with Heartsome Translation Studio
- CSV, TXT parallel texts to TMX Tms with Heartsome
- Okapi Rainbow to PO, TMX, CSV
- SDL/Trados TM and TB to TMX with TradosStudio resources converter



Overview: part II

typical tasks of a translator

- ✓ translate a website
- ✓ create a TM on the basis of existing translations
- ✓ manage terminology and dictionaries
- ✓ extract terminology from texts
- ✓ use machine translation
- ✓ convert file formats
- ✓ **manage bilingual files**
- ✓ manage pdf files
- ✓ quality assurance
- ✓ use text corpora



► bilingual file types

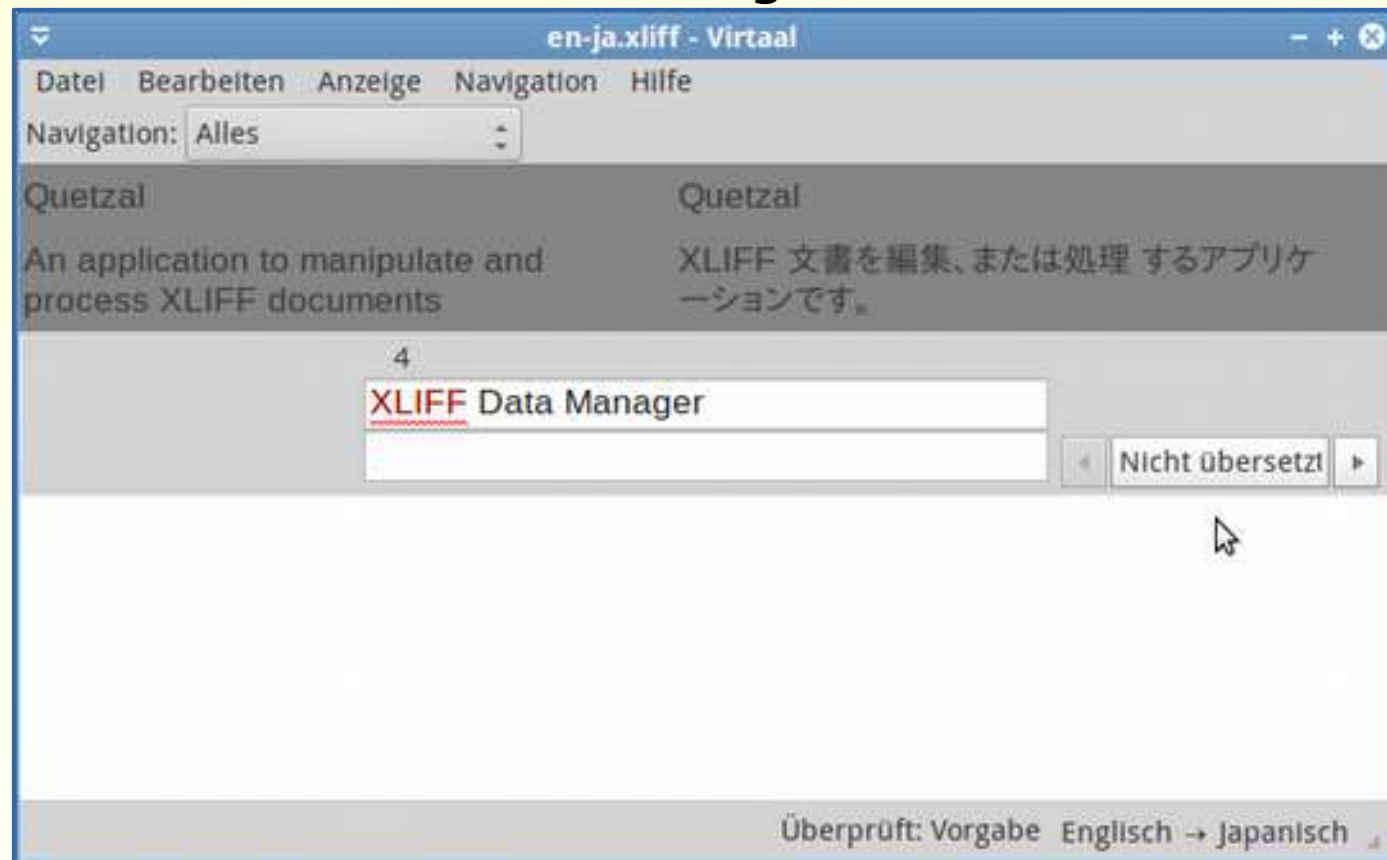
text files which include source text segments aligned to target text segments, i.e. files which contain translations for all segments or just for a part of segments

XLIFF, sdlxliff, ttx

can not be translated with OmegaT directly, or rather only when *source text segment = target text segment*

existing translations can thus not be leveraged

*partly translated
file as shown
in Virtaal*



▶ bilingual file types

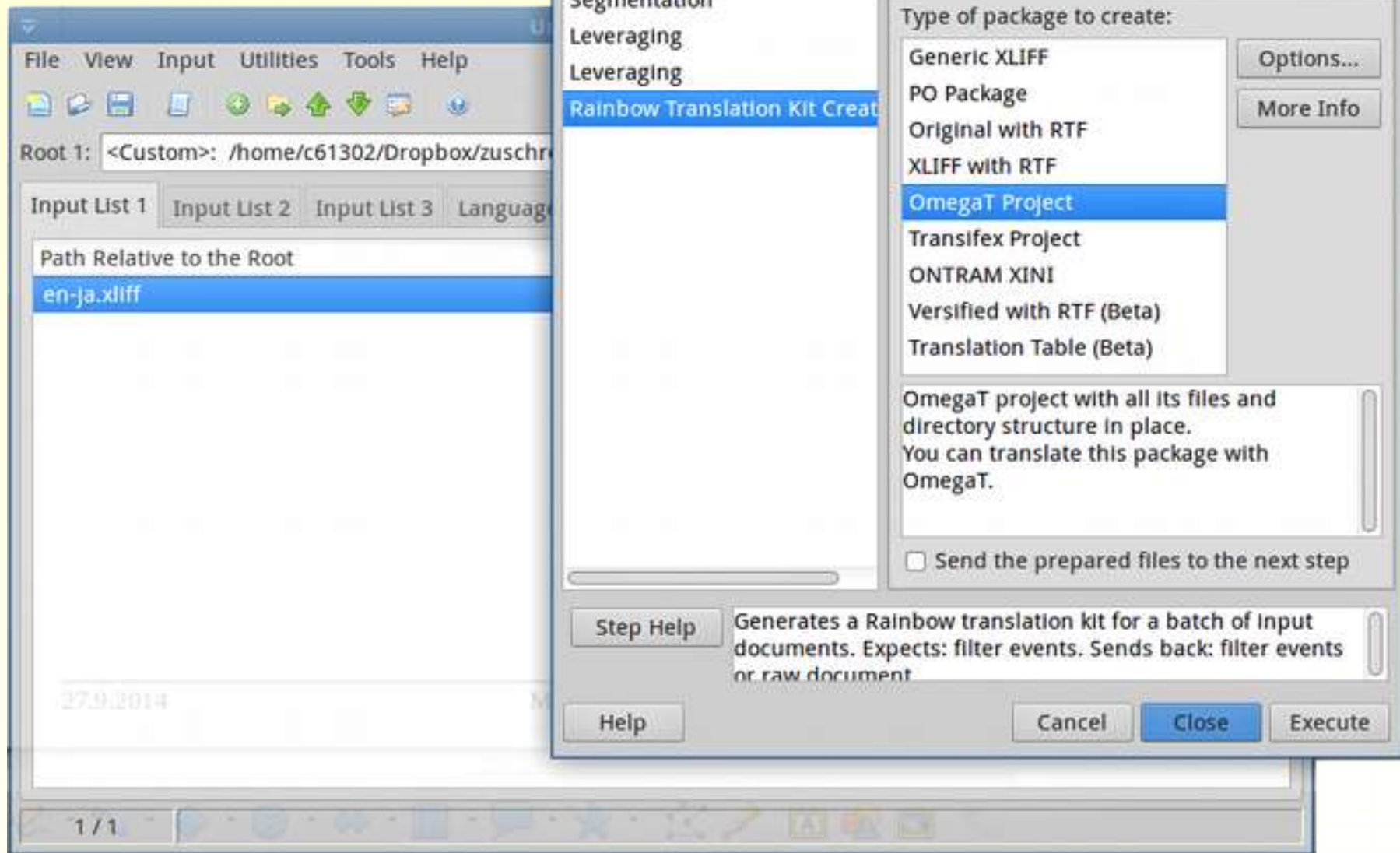


Existing translations can be re-used in OmegaT by using Okapi Rainbow and by creating ready-to-go OmegaT translation projects

- XLIFF
 - sdxliff
 - ttx
- }
- extract source text
 - create TM from existing translations
 - translate project in OmegaT
 - post process translation in Rainbow
 - get translated bilingual files

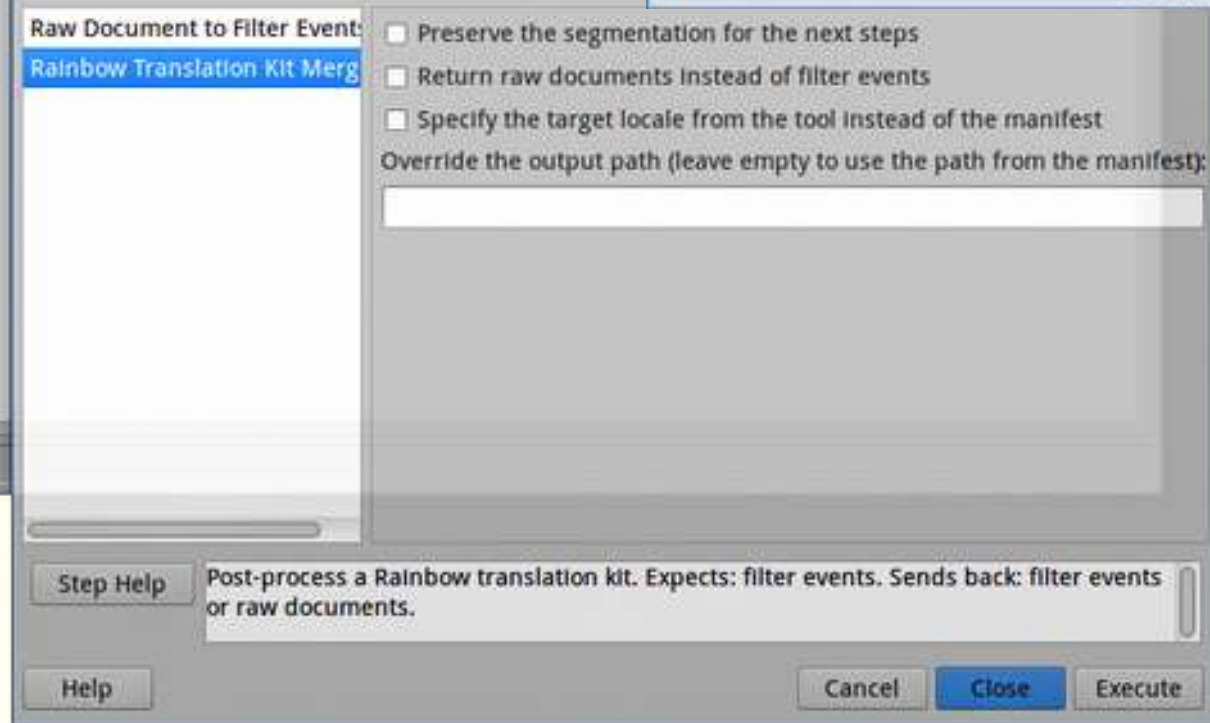
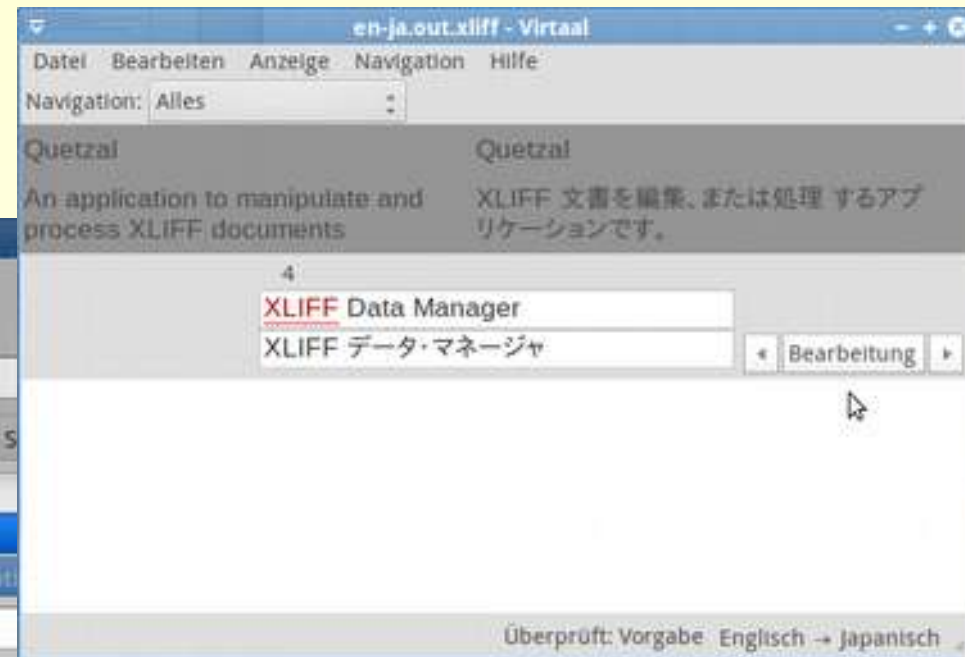
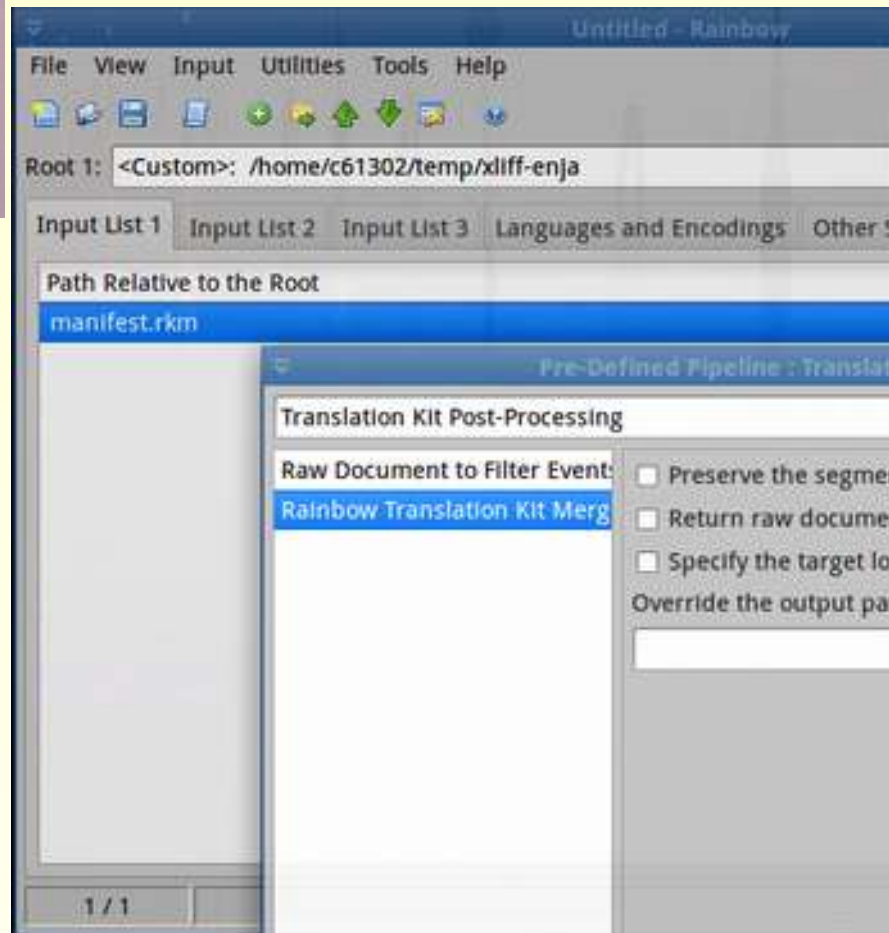
Okapi Rainbow

translation kit creation



Okapi Rainbow

translation kit post processing



Overview: part II

typical tasks of a translator

- ✓ translate a website
- ✓ create a TM on the basis of existing translations
- ✓ manage terminology and dictionaries
- ✓ extract terminology from texts
- ✓ use machine translation
- ✓ convert file formats
- ✓ manage bilingual files
- ✓ **manage pdf files**
- ✓ quality assurance
- ✓ use text corpora



► PDF files

- translate PDF files with OmegaT
only text-based PDF and target text will be txt file
- extract text from PDF files
with gPDFText eBook-Editor (Linux)
- split and merge PDF files
with PDF-Sam
- annotate PDF files
with Xournal (Linux)
- compare PDF files with DiffPDF



PDF and OmegaT

- translate text-based PDF directly in OmegaT



A screenshot of the OmegaT 3.14.1 software interface. The main window shows a PDF document titled "9th PCTS Conference - 1st call EN.pdf". The text in the document is highlighted in green. A "Fuzzy Matches" window is open on the right, showing a list of matches with their respective percentages and file paths. A "Project Files (3)" dialog box is open in the foreground, displaying a table of files and their statistics. The table has columns for "Filename", "Filter", "Encoding", "Number ...", and "Number ...". The files listed are "9th PCTS Conference - 1...", "9th PCTS Conference - R...", and "PCTS9 Program+Book of ...". The dialog box also has buttons for "Move First", "Move Up", "Move Down", "Move Last", "Copy Files to Source Folder...", "Download MediaWiki Page...", and "Close". The main window also shows a "Call for Papers" section and a "Conference sections: Str." section. The bottom status bar shows "0/26 (0/890, 895) 114/0".

extract text from PDF

- gPDFText eBook-Editor extracts Text from text-based PDF without line breaks

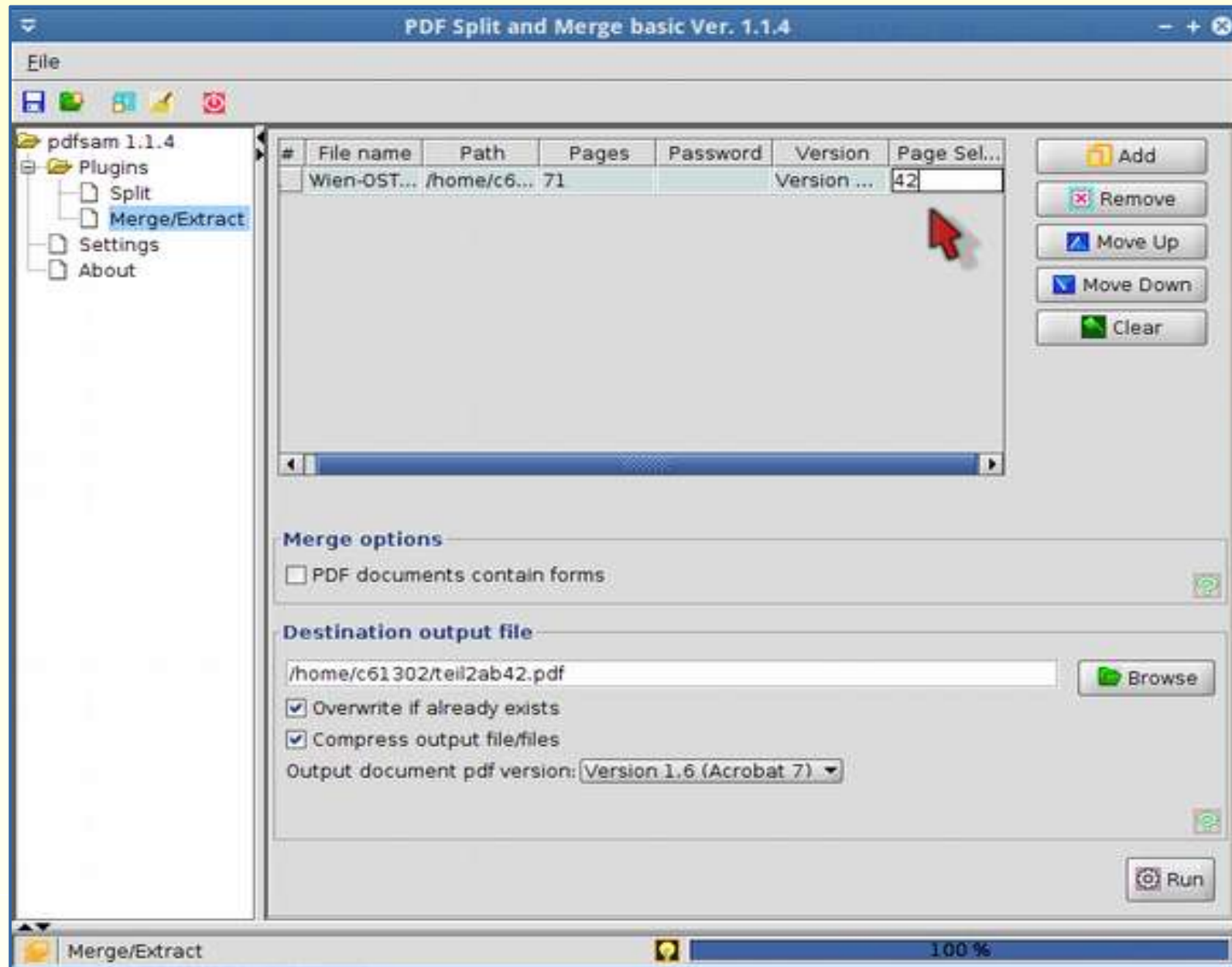
The image displays the gPDFText eBook-Editor interface. The main window shows the extracted text from a PDF, which is displayed as a single continuous block of text without line breaks. The text includes contact information for the 9th International Conference on Professional Communication and Translation Studies, held from March 26-27, 2015, at Politehnica University of Timisoara. The text also lists conference sections such as Communication and public relations, Linguistics, Translation studies, and Foreign language teaching. Additionally, it provides details about the types of presentations, working languages, and publication information.

A settings dialog box titled "gPDFText-Einstellungen" is open, showing options for standard paper format (A4-Papierformat (Portrait) is selected), settings for regular expressions (Einzelne Zeilen zusammenfügen and mit Bindestrich getrennte Wörter vereinen are checked), and font settings (Standard-Schriftart: Sans, 12). The Lexikon field is empty.

The status bar at the bottom indicates "Page 1 of 1", "899 words, 1.381 characters", and "Default Style: German (Germany)". A notification at the bottom right states "Text-Datei gespeichert." (Text file saved).

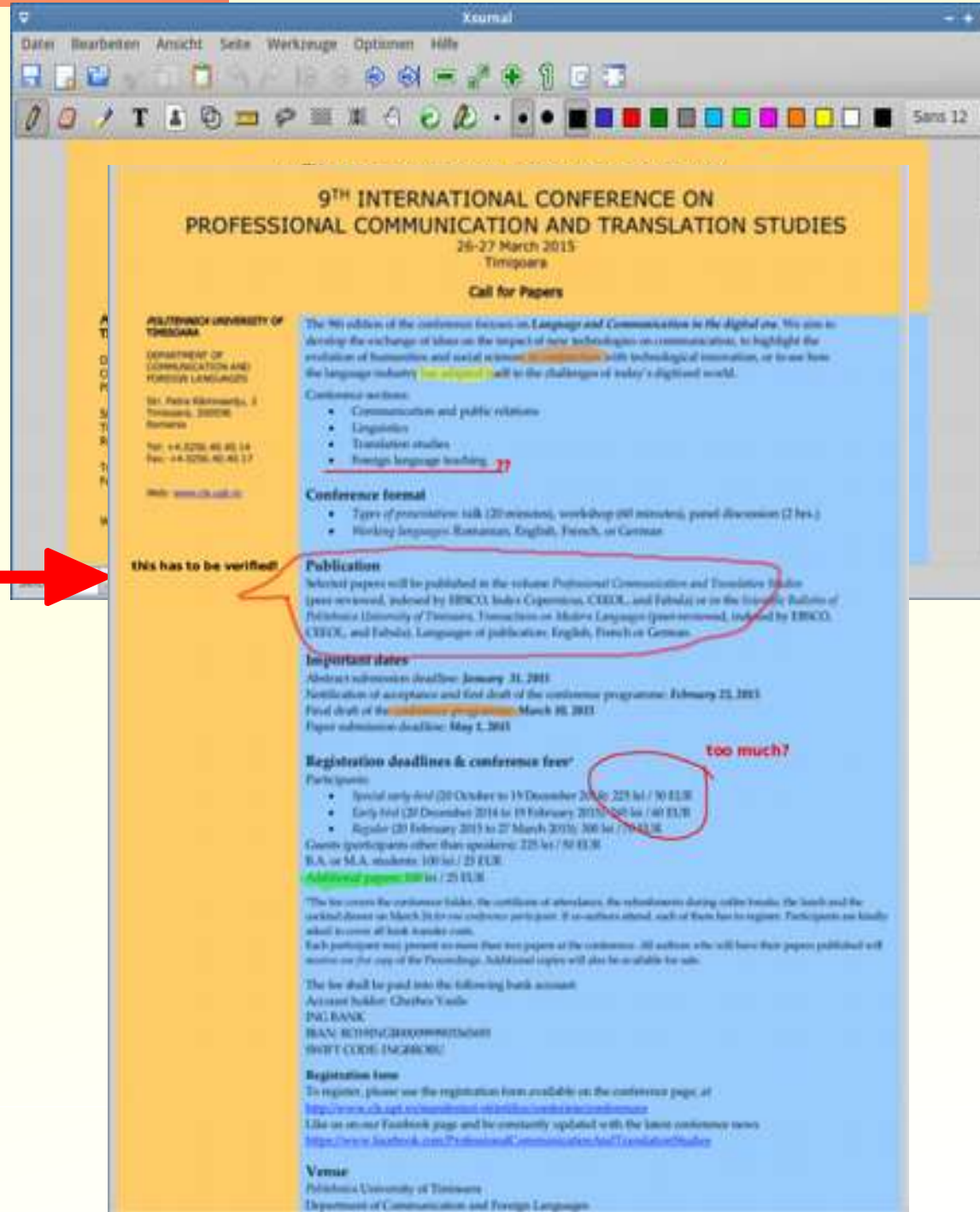
split and merge PDFs

- PDF-Sam



PDF-Dateien annotieren

- comment, underline, highlight, etc. with Xournal



Overview: part II

typical tasks of a translator

- ✓ translate a website
- ✓ create a TM on the basis of existing translations
- ✓ manage terminology and dictionaries
- ✓ extract terminology from texts
- ✓ use machine translation
- ✓ convert file formats
- ✓ manage bilingual files
- ✓ manage pdf files
- ✓ **quality assurance**
- ✓ use text corpora



▶ Quality assurance

- 1) OmegaT QA scripts
- 2) Okapi Checkmate
- 3) TMX-Validator
- 4) XLIFF Checker



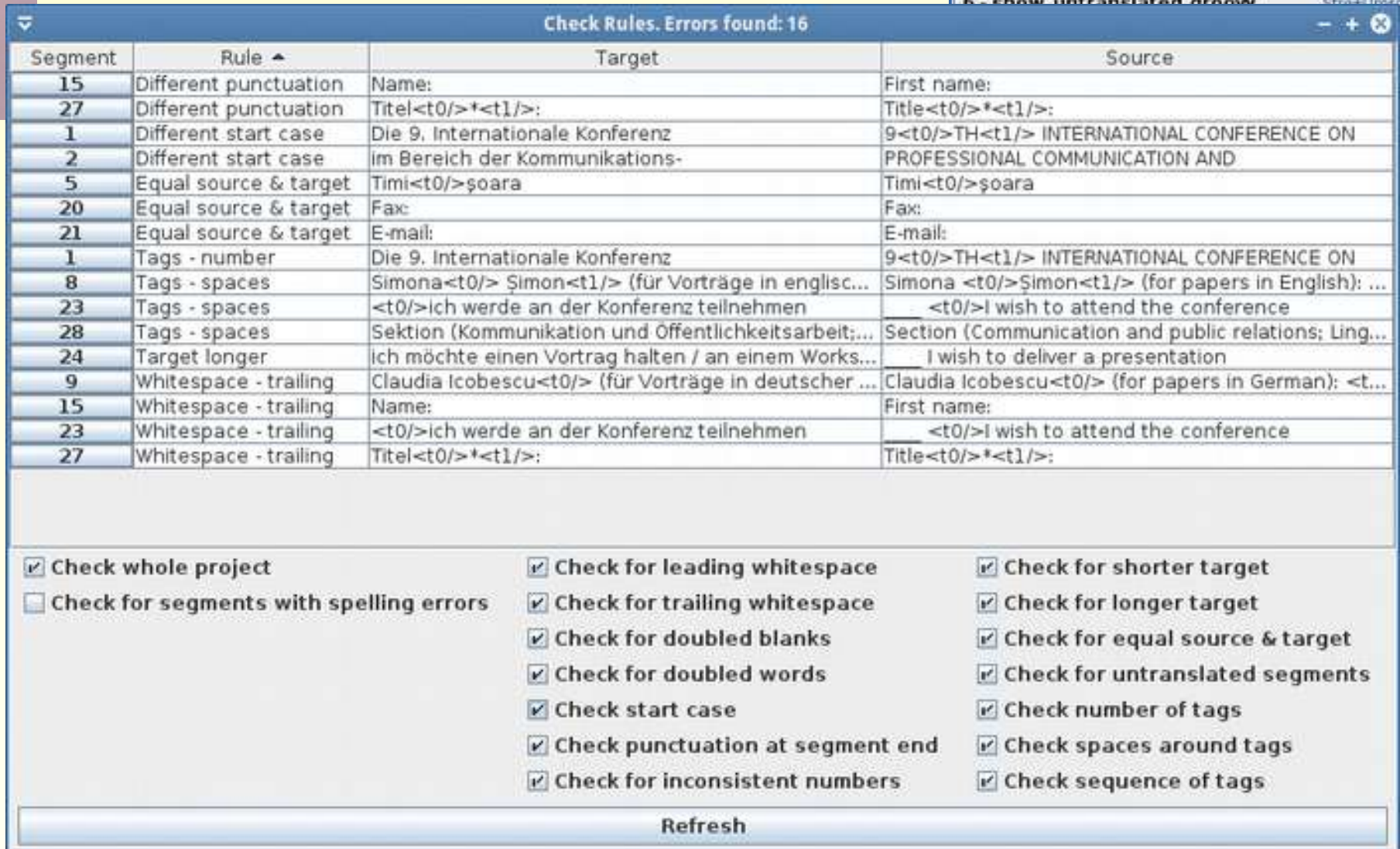
QA with OmegaT

- OmegaT QA scripts



Extras Optionen Folders Hilfe

- Tags überprüfen Strg+Umschalt-V
- Validate Tags for Current Document
- Statistiken
- Treffer-Statistiken
- Match Statistics per File
- Skripting...
- 1 - Example - Search and Replace Strg+Umschalt-F1
- 2 - QA - Check Rules** Strg+Umschalt-F2
- 3 - Strip Tags Strg+Umschalt-F3
- 4 - Spellcheck Strg+Umschalt-F4
- 5 - QA - Identical Segments Strg+Umschalt-F5
- 6 - show untranslated groups Strg+Umschalt-F6



Check Rules. Errors found: 16

Segment	Rule	Target	Source
15	Different punctuation	Name:	First name:
27	Different punctuation	Titel<t0/>*<t1/>:	Title<t0/>*<t1/>:
1	Different start case	Die 9. Internationale Konferenz	9<t0/>TH<t1/> INTERNATIONAL CONFERENCE ON
2	Different start case	im Bereich der Kommunikations-	PROFESSIONAL COMMUNICATION AND
5	Equal source & target	Timi<t0/>șoara	Timi<t0/>șoara
20	Equal source & target	Fax:	Fax:
21	Equal source & target	E-mail:	E-mail:
1	Tags - number	Die 9. Internationale Konferenz	9<t0/>TH<t1/> INTERNATIONAL CONFERENCE ON
8	Tags - spaces	Simona<t0/> Simon<t1/> (für Vorträge in englisc...	Simona <t0/>Simon<t1/> (for papers in English): ...
23	Tags - spaces	<t0/>ich werde an der Konferenz teilnehmen	<t0/>I wish to attend the conference
28	Tags - spaces	Sektion (Kommunikation und Öffentlichkeitsarbeit;...	Section (Communication and public relations; Ling...
24	Target longer	ich möchte einen Vortrag halten / an einem Works...	I wish to deliver a presentation
9	Whitespace - trailing	Claudia Icobescu<t0/> (für Vorträge in deutscher ...	Claudia Icobescu<t0/> (for papers in German): <t...
15	Whitespace - trailing	Name:	First name:
23	Whitespace - trailing	<t0/>ich werde an der Konferenz teilnehmen	<t0/>I wish to attend the conference
27	Whitespace - trailing	Titel<t0/>*<t1/>:	Title<t0/>*<t1/>:

Check whole project

Check for segments with spelling errors

Check for leading whitespace

Check for trailing whitespace

Check for doubled blanks

Check for doubled words

Check start case

Check punctuation at segment end

Check for inconsistent numbers

Check for shorter target

Check for longer target

Check for equal source & target

Check for untranslated segments

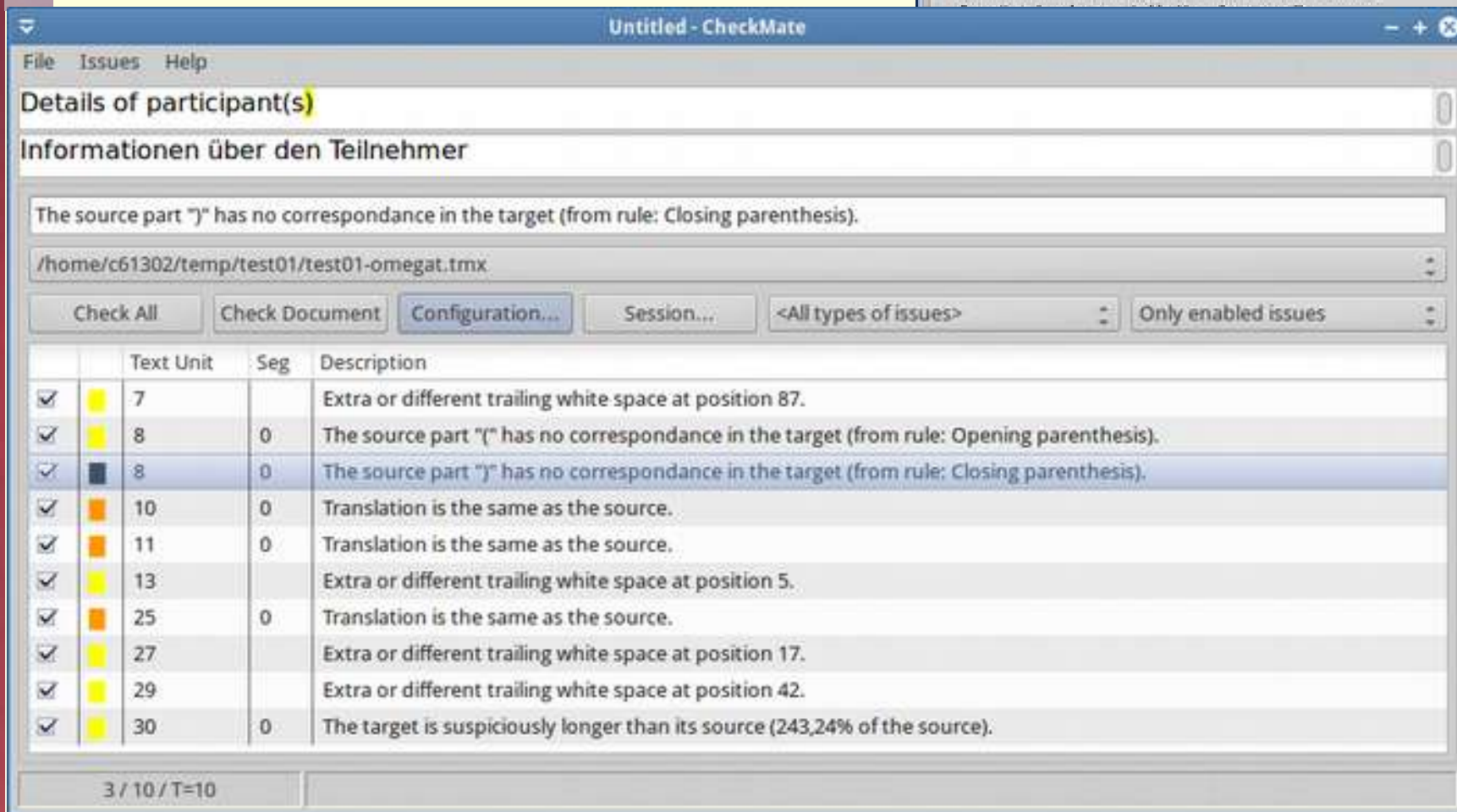
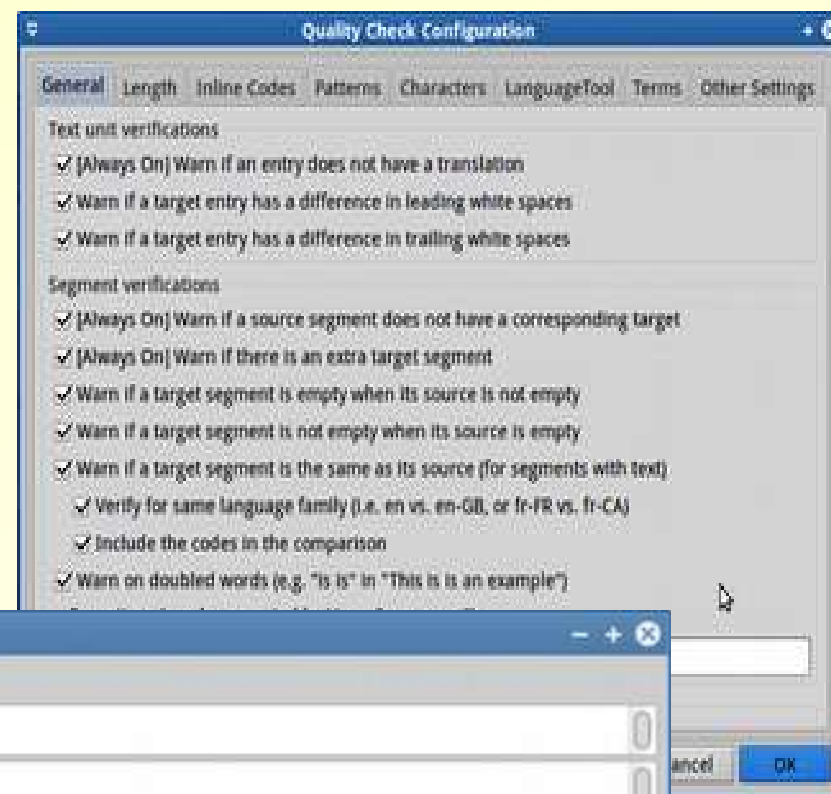
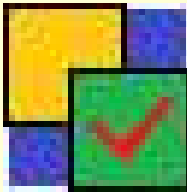
Check number of tags

Check spaces around tags

Check sequence of tags

Refresh

QA with Checkmate

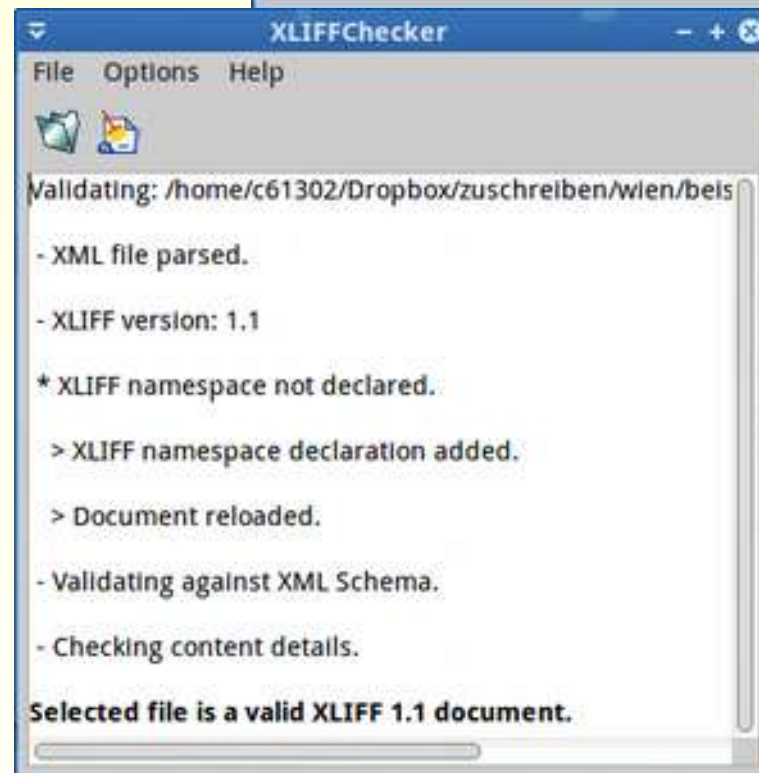


formal validation

TMX-Validator



XLIFF-Checker



Overview: part II

typical tasks of a translator

- ✓ translate a website
- ✓ create a TM on the basis of existing translations
- ✓ manage terminology and dictionaries
- ✓ extract terminology from texts
- ✓ use machine translation
- ✓ convert file formats
- ✓ manage bilingual files
- ✓ manage pdf files
- ✓ quality assurance
- ✓ **use text corpora**



▶ create a reference corpus

Why?

search for terms, collocations, idioms ...

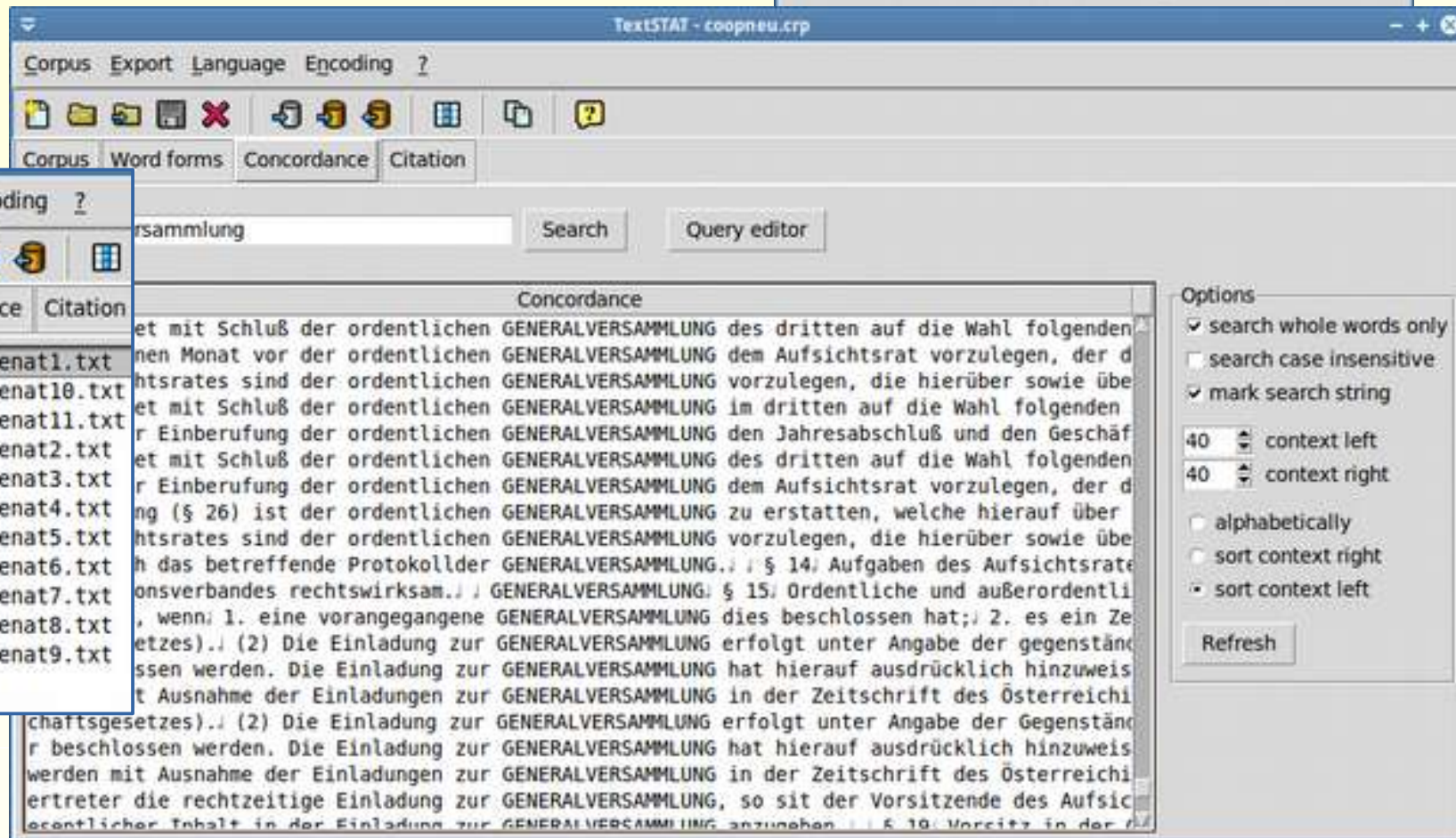
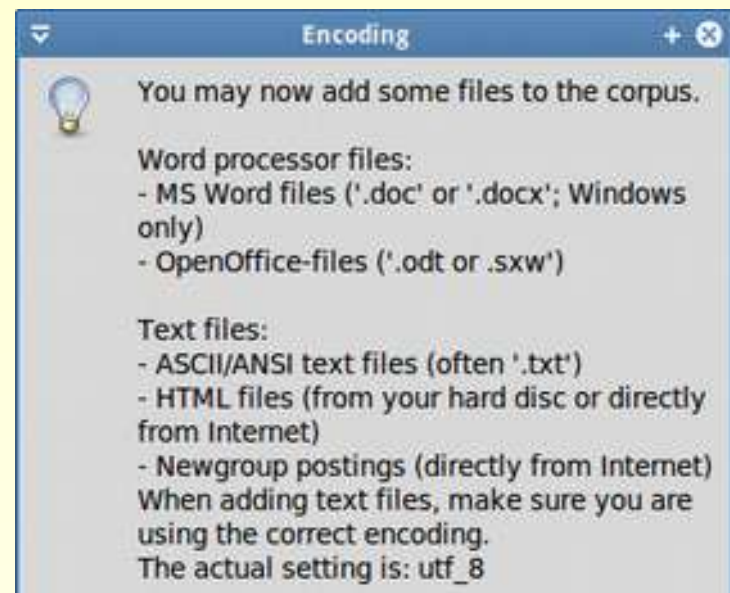
what you need:

- a representative quantity of LSP texts in the subject domain you want to translate
- a free concordance tool



TextSTAT

- file formats:
doc, odt, html, txt
- concordance search
- frequency lists



AntConc

- AntConc

The screenshot displays the AntConc 3.4.1u (Linux OS) 2014 interface. The main window shows a list of corpus files on the left, including COOP1.TXT through COOP15.TXT and COST45.TXT. The central pane displays a concordance table for the search term 'assemblea ordinaria'. The table has columns for Rank, Freq, Range, and Cluster. The search settings at the bottom are configured for 'Words', 'Case', and 'Regex' options, with a search window size of 50. The search results are displayed in a color-coded format, highlighting the search term and its context.

Rank	Freq	Range	Cluster
10	6	6	assemblea ordinaria indetta dal consiglio stesso.
11	6	6	assemblea ordinaria è
12	6	6	assemblea ordinaria: 1) approva il bilancio cons
13	6	6	assemblea ordinaria: 1) approva il bilancio cons
14	4	4	assemblea ordinaria deve
15	4	4	assemblea ordinaria deve essere
16	4	4	assemblea ordinaria deve essere convocata
17	3	3	assemblea ordinaria deve essere convocata almen
18	3	3	assemblea ordinaria deve essere convocata almen
19	3	3	assemblea ordinaria e straordinaria può
20	3	3	assemblea ordinaria e straordinaria può essere
21	3	3	assemblea ordinaria è validamente
22	3	3	assemblea ordinaria è validamente costituita
23	3	3	assemblea ordinaria è validamente costituita con

Search Term: assemblea ordinaria
Cluster Size: Min. 3, Max. 7
Search Window Size: 50
Kwic Sort: Level 1 LR, Level 2 2R, Level 3 3R

less file formats
more linguistic
analysis

Overview: part II

typical tasks of a translator

- ✓ translate a website
- ✓ create a TM on the basis of existing translations
- ✓ manage terminology and dictionaries
- ✓ extract terminology from texts
- ✓ use machine translation
- ✓ convert file formats
- ✓ manage bilingual files
- ✓ manage pdf files
- ✓ quality assurance
- ✓ use text corpora

... + ...

translation of subtitles

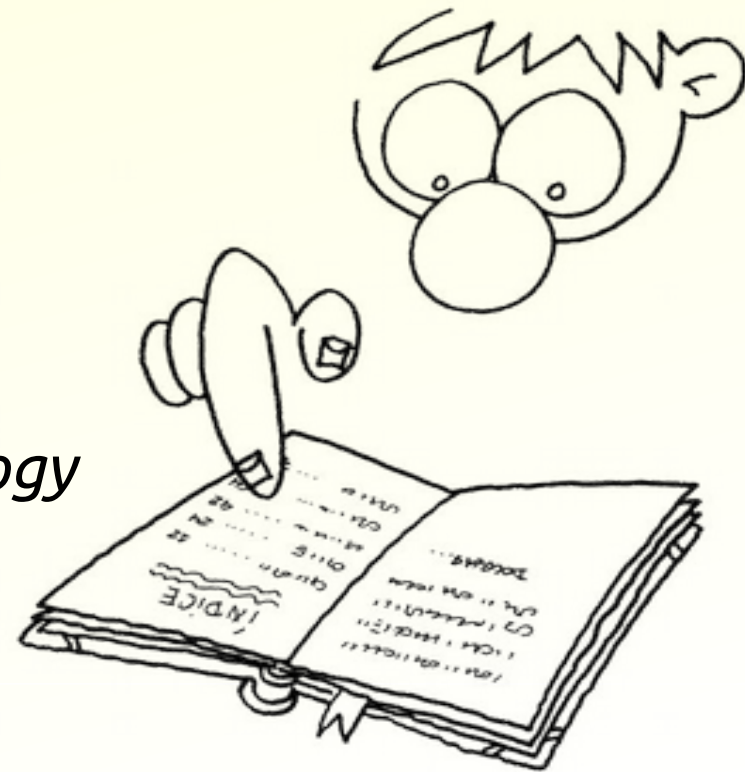
editing and translating PO files

creating mindmaps for terminology

creating a web corpus

localizing software

evaluating MT output ...



What next?

Simply, try it out!

*Thank you
for your attention!*



<http://www.petersandrini.net>

<http://uibk.academia.edu/PeterSandrini>