Using European Translation Tools with Asian Languages

Introduction

- Degree in translation (Chinese/Japanese into German), Computational Linguistics
- Worked for Trados in Japan, Germany, USA
- Since 2000, independent trainer and consultant for translation tools (TM tools and terminology management tools)
- 2 employees for technical support and terminology management

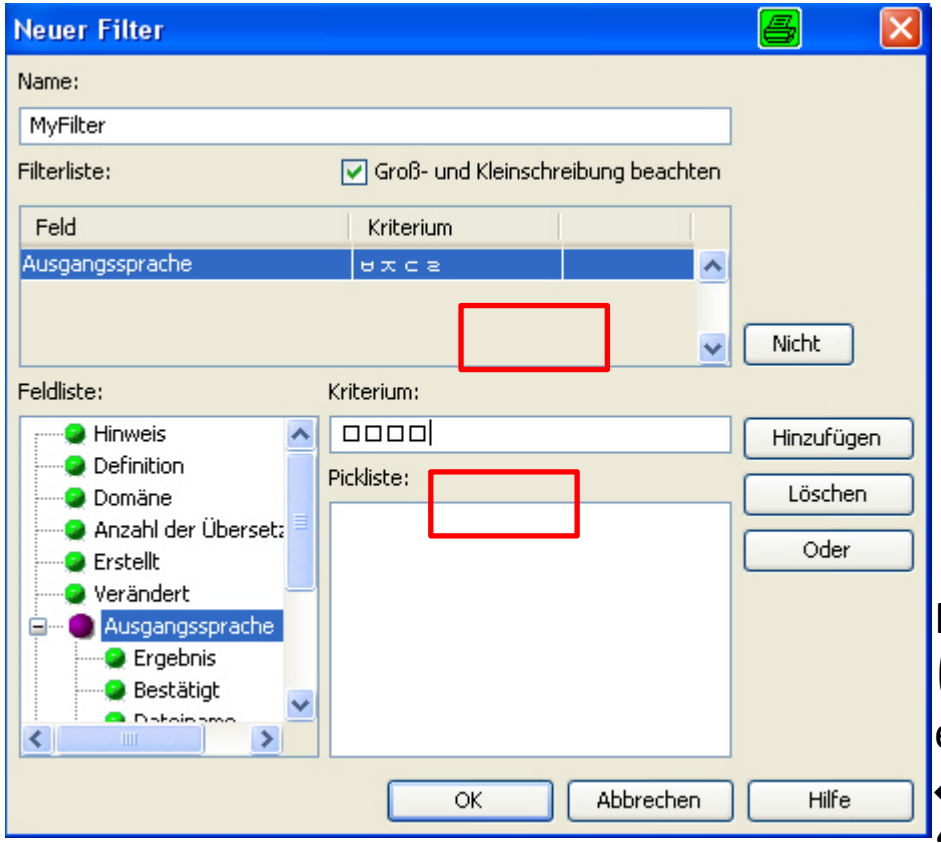
Agenda

- History ☺
- User Interface
- Display of Asian Characters
- Terminology management systems
- QA issues/possibilities
- Segment matching and sub-segment matching
- Statistics

History

- When I started with Trados tools in 1997, Trados was just entering the Japanese market.
- One of the main issues was the inability of the tool to take Japanese characters in the names of user-defined fields and Mojibake -> character corruption in the translation memory database, mostly because of missing Unicode support
- For some features to work correctly, a localized operating system was needed

History



文字化け

•¶Žš‰»,¯,

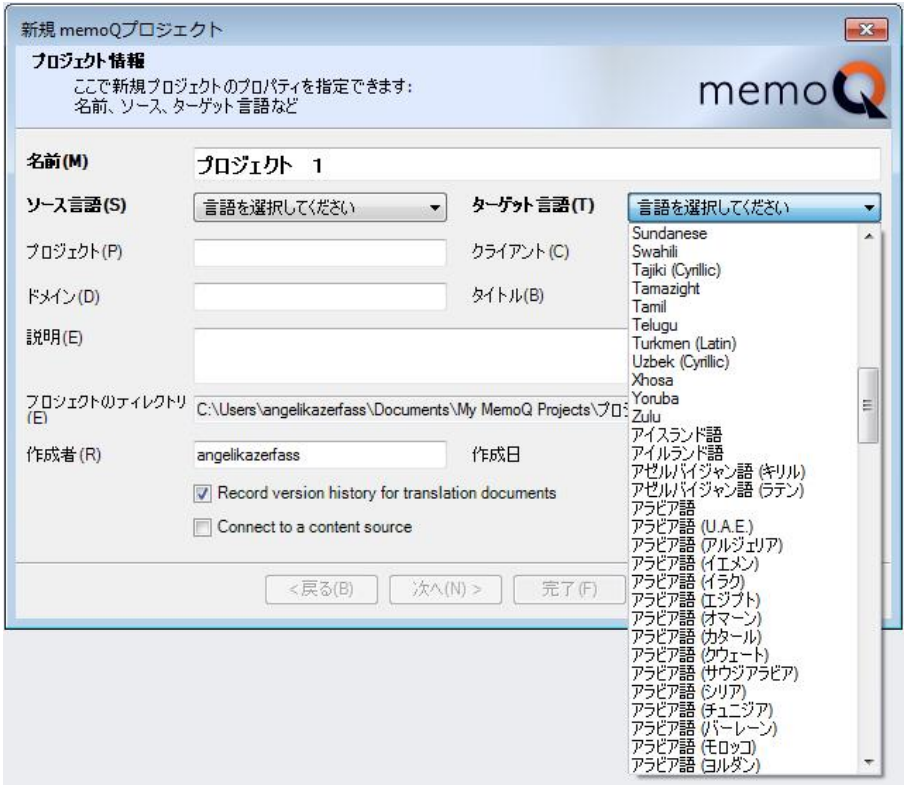
For example The set of bytes that says “文字化け” (the word for “mojibake” in Japanese) encoded in UTF-8 would show up as “��絖�����” in EUC-JP, “������������” in ISO-2022-JP, and “æ–‡å—åŒ–ã‘” in ISO-8859-1.

Today

- Character corruption does not happen as often any more, but still there are some places where tools will not be able to take Asian characters or where functionalities that work well with latin-based scripts will not work with Asian texts, like terminology extraction, sub-segment matching, etc.

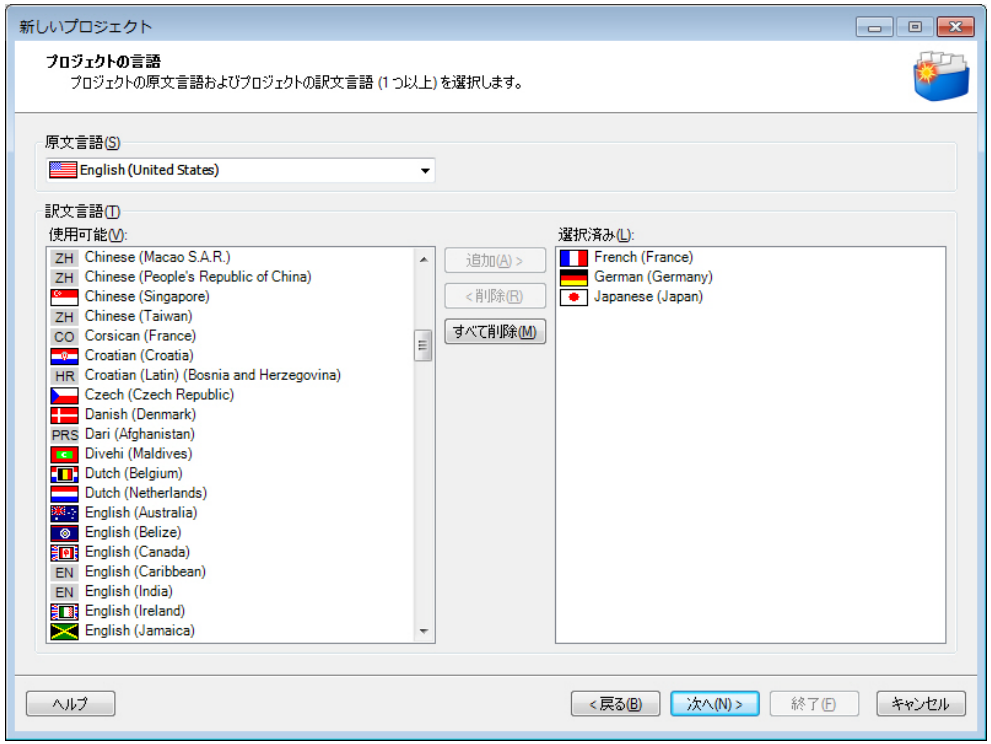
User Interface

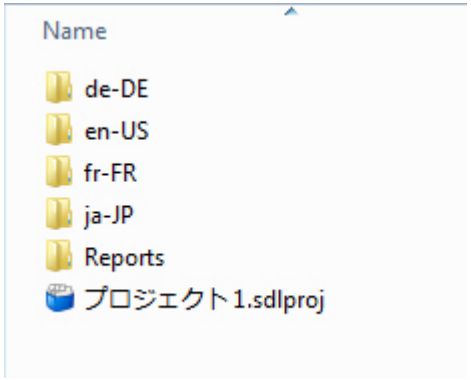
- It is still the case that the user interface of a translation tool might only be available in English or that the user interface does not get translated as quickly for Asian languages as for other languages.



User Interface

- There might also be decisions of the tools vendors to leave ceratain things in English for all languages (names of languages, names of folders).





User Interface

- An issue for some users is the default font used for an Asian language as this might be right for the Asian characters, but might not look good for latin characters.

Sample File	サンペルファイル(sample file)
Table of Contents	

Using Asian characters in project names…

- Whereas in earlier tools Asian characters could often not be used to name projects, files or user-defined fields, this is now possible almost everywhere.
- Mixing display languages is also not an issue any more (the project is named in Japanese, the user interface is set to Chinese).

Terminology Management

- Term bases that are set up for European languages often do not work for Asian languages
  - European: single words are terms (system)

Terminology Management

- Asian languages are more contextual. The term "system" can be translated with many different characters in Chinese, depending on the kind of system…

system {Substantiv}

system {Subst.} (auch: scheme, tract, economy, systems)	系统 [xìtǒng] {Subst.}
HEALS (Honeywell Error Analysis and Logging **System**)	霍尼韦尔错误分析与记录**系统**(Honeywell公司)
system {Subst.} (auch: scheme, systems)	体系 [tǐxì] {Subst.}
agricultural resources and ecological protection **system**	农业资源和生态环境保护**体系**
system {Subst.} (auch: institution, regime, systems)	制度 [zhìdù] {Subst.}
CAMELS rating **system**	CAMELS评级**制度**
system {Subst.} (auch: scheme, systems)	体系 [tǐxì] {Subst.}
system {Subst.} (auch: institution, regime, systems)	制度 [zhìdù] {Subst.}
system {Subst.} (auch: framework, organization, systems)	体制 [tǐzhì] {Subst.}
the Central Committee of CPC's decision on economic **system** reform	中共中央关于经济**体制**改革的决定
system {Subst.} (auch: method, order, systems, orders)	秩序 [zhìxù] {Subst.}
system {Subst.} (auch: law, rule, systems, laws)	规律 [guīlǜ] {Subst.}
system {Subst.} (auch: approach, device, measures, medium)	方法 [fāngfǎ] {Subst.}
system {Subst.} (auch: framework, organization, systems)	体制 [tǐzhì] {Subst.}
system {Subst.} (auch: approach, device, measures, medium)	方法 [fāngfǎ] {Subst.}

Terminology Management

- Term components of translation memory tools often offer fuzzy matching (like with segments from the translation memory) or matching of word stems (i.e. ignoring term endings during comparison).

Terminology Extraction

- Term extraction tools can be split into two groups:
  - Statistical extraction tools
  - Linguistic extraction tools
- Most extraction tools available at the moment are geared towards European languages, although, in theory, statistical tools could also extract fom Asian texts.
- Especially for Japanese and Chinese, where thre are no spaces to delimit words, the term extaction tools do not work well.

This is, how a text looks to a statistical extraction tool…

Vot gnig harengoga fuor tok gnig nor shewerginhatz. Mirhon bortup tip trewshu gnig batbo loqtet. Bortup ter, bortup nofdas, semsel nih furpo ayano bliktreptat. Mirhon granbevtrov driktopret grig go wasbrekit mut mirkep taptro gnig suf. Aktrep zitpek nitnit bortup mil. Setrimb ak troptan bur metlatkento.

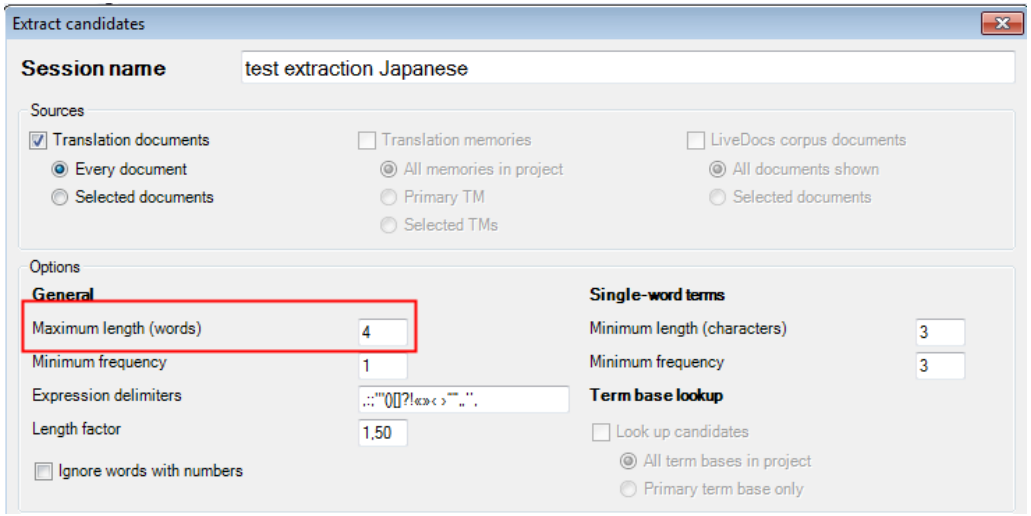
Term candidates – statistical extraction

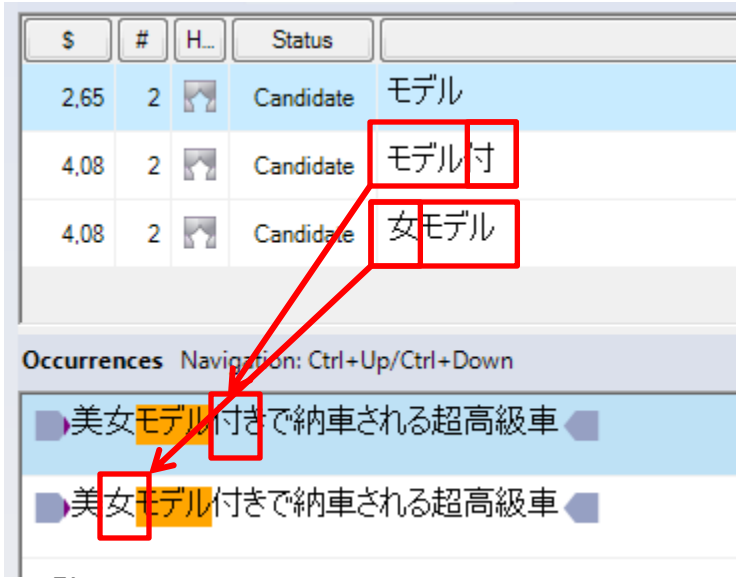
- Statistical extraction
  - Monolingual or bilingual
  - Suitable for every language / language combination (for example from a translation memory)
  - The larger the collection of extraction material, the better the extracted lists
  - Stop word lists
  - Context sentences
  - In theory for all languages, but in praxis Asian languages need more selection work than others

Words	Freq	English (United States)
4	2	high quality product carries
5	2	high quality product carries an
2	1	High speed
2	2	high speed
3	1	High speed cooling
4	1	High speed cooling fan
5	1	High speed cooling fan relay
3	1	high speed engine
4	1	high speed engine cooling
5	1	high speed engine cooling fan
2	1	high temperatures

Term candidates – statistical extraction

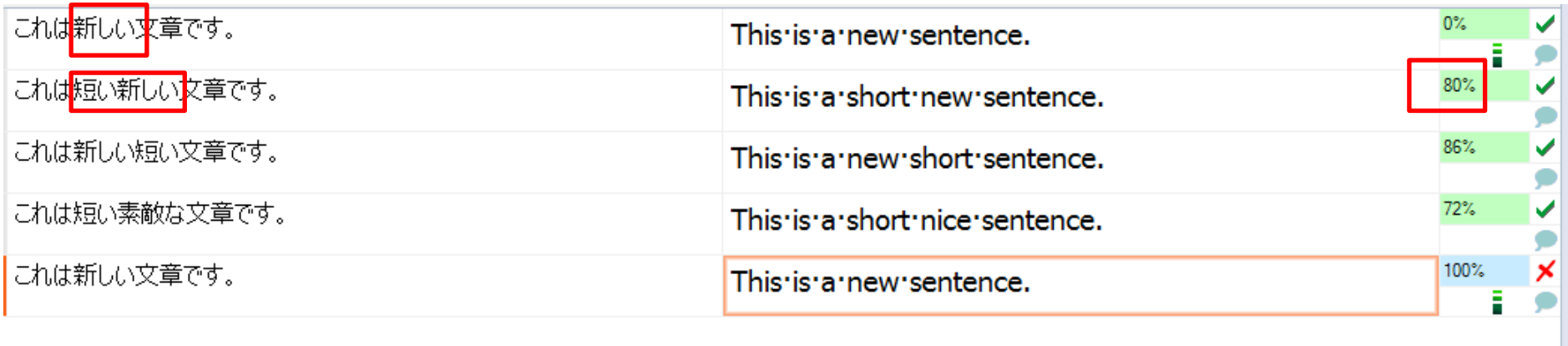
- Statistical extraction
  - Settings are focused on words surrounded by spaces as delimiters
  - Result: "stupid" term candidates





Match Values

- Different matching values because of algorithms (taking into account number of words and spaces and the words themselves in European languages, taking into account number of characters and characters themselves)



Using Asian characters in project names…

- In file names
- In project names
- In regex
- When translating into Asian languages (fonts/sizes used)

Term Management

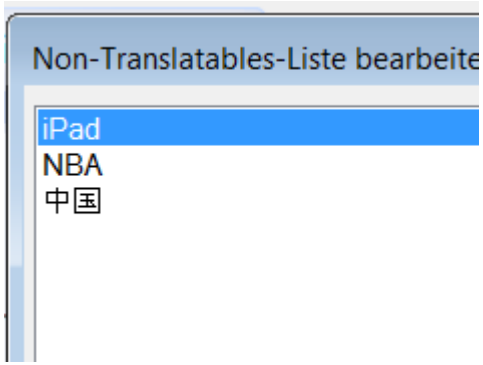
- Term extraction methods / settings
- Term representation in term base
- Term recognition during translation / settings
  - Input editor does not allow predictive typing (memoQ) or AutoSuggest (Trados) for Chinese, Japanese. But AutoPick and results from list in memoQ help. Studio does not offer takeover of terms other than starting to type
- Term checking (rate of error messages)

QA possibilities/issues

- Counting length of text (double-byte characters as 2?)
- Regex using hex codes does not work with Japanese in memoQ (see conversion tool http://weber.ucsd.edu/~dkjordan/resources/unicodemaker.html)

Issues memoQ

- Non-trans list
  - Term input from translation doc will show up in gray in results list (NBA)



  - Term typed in using input method editor chinese, does not get recognized (NBA)