Terminology post-editing of neural MT by UTX glossary data

MT Summit 2017

YAMAMOTO Yuji
UTX team leader, AAMT
http://www.aamt.info/english/utx/

Presenter: Yamamoto Yuji

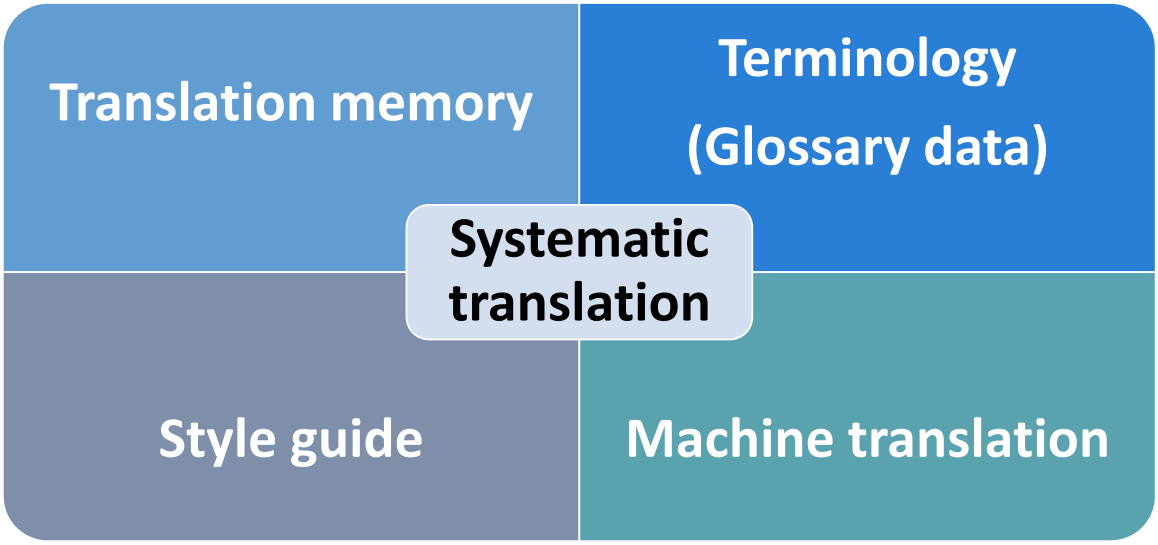
- CosmosHouse Founder/Representative
- Language/translation consultant
- AAMT UTX team leader
- ISO/TC37 (terminology) committee member
- Contact at http://cosmoshouse.com/mail.htm

Agenda

1. Background – terminology and NMT
2. UTX – a structured glossary format
3. Terminology post-editing
4. Conclusion

Background – terminology and NMT

Terminology is an essential part of systematic translation



Commercial translation requirements

■Glossaries are established by client companies.

e.g. Microsoft glossaries

■Use of a company vocabulary is not optional. You are required to use certain terms for translation.

NMT problems

1. NMT mistranslates low-frequency words
2. NMT cannot reflect an existing glossary
3. NMT lacks terminological consistency

Problem 1: NMT mistranslates less-frequent terms

- such as proper names and technical terms
  - e.g. auxiliary verb→*補助動詞 (助動詞 is correct)
  - response rate→*奏功率 (回答率 is correct)
- Missing or repeated translation

Problem 2: NMT cannot reflect an existing glossary

- e.g. “liaison” in ISO context
- A glossary is not an issue for general MT users
- A glossary is essential in a systematic translation
- Many companies are not managing glossaries in an organized manner
- Translation problems are hidden in such an environment

Problem 3: NMT lacks terminological consistency

- e.g. International Standard→国際規格、国際標準
  - resource→資源、リソース
- Terminology consistency is not an issue for general MT users
- But terminology consistency is important in systematic translation

Prevalence of RbMT in Japan

- Strong demands for translation.
- EN-JA bilingual market.
- Early MT commercialization since 1990s.
- Many commercial RbMT packages are sold.

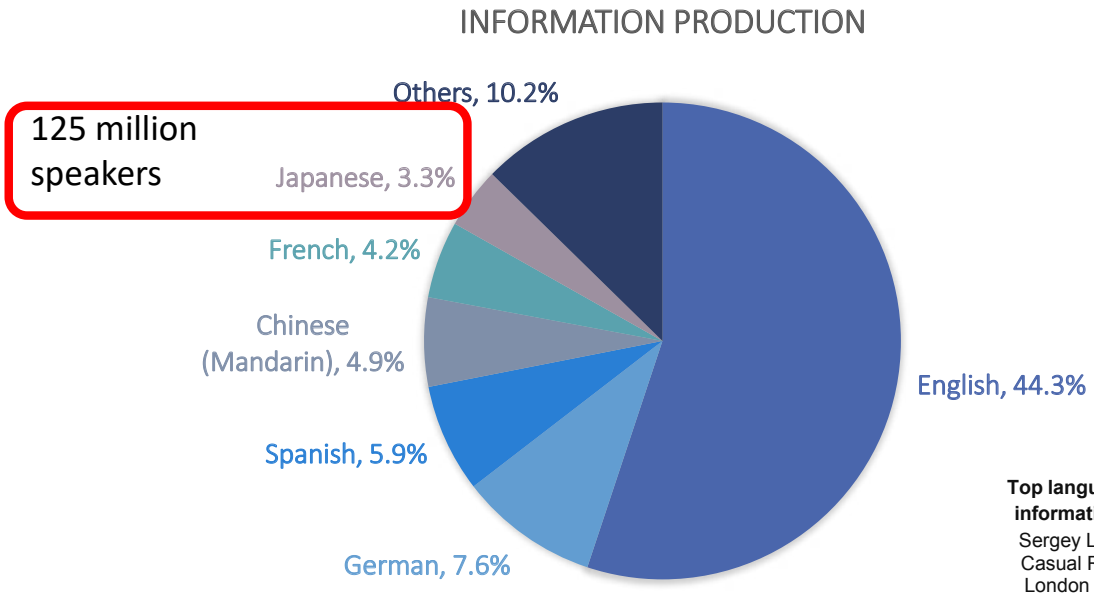
Toshiba

Fujitsu

Cross Language

Kodensha

Japanese is an influential language, but its market is bilingual



Top languages in global information production
Sergey Lobachev
Casual Reference Librarian
London Public Library

https://journal.lib.uoguelph.ca/index.php/perj/article/view/826/1358#.WY_eh1GrSHs

Bilingual or multilingual scenario?

- Japan – Japanese and English
- Europe, Americas, Africa, etc. - multilingual

Terminology management must be simplified

- Or it will never be implemented.
- Multilingual complexity is not necessary for a bilingual environment.
- A simple Excel sheet is too simple.

UTX – a structured glossary format

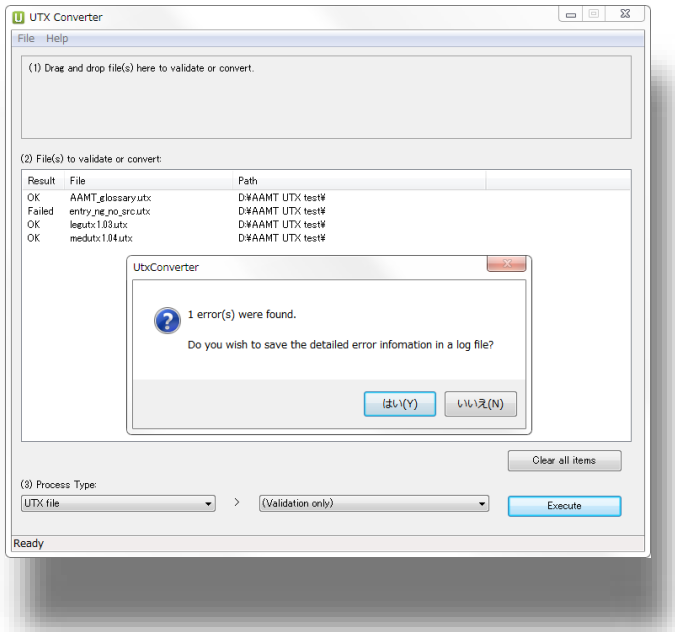
What is UTX (Universal Terminological eXchange)?

Simple but structured glossary data format

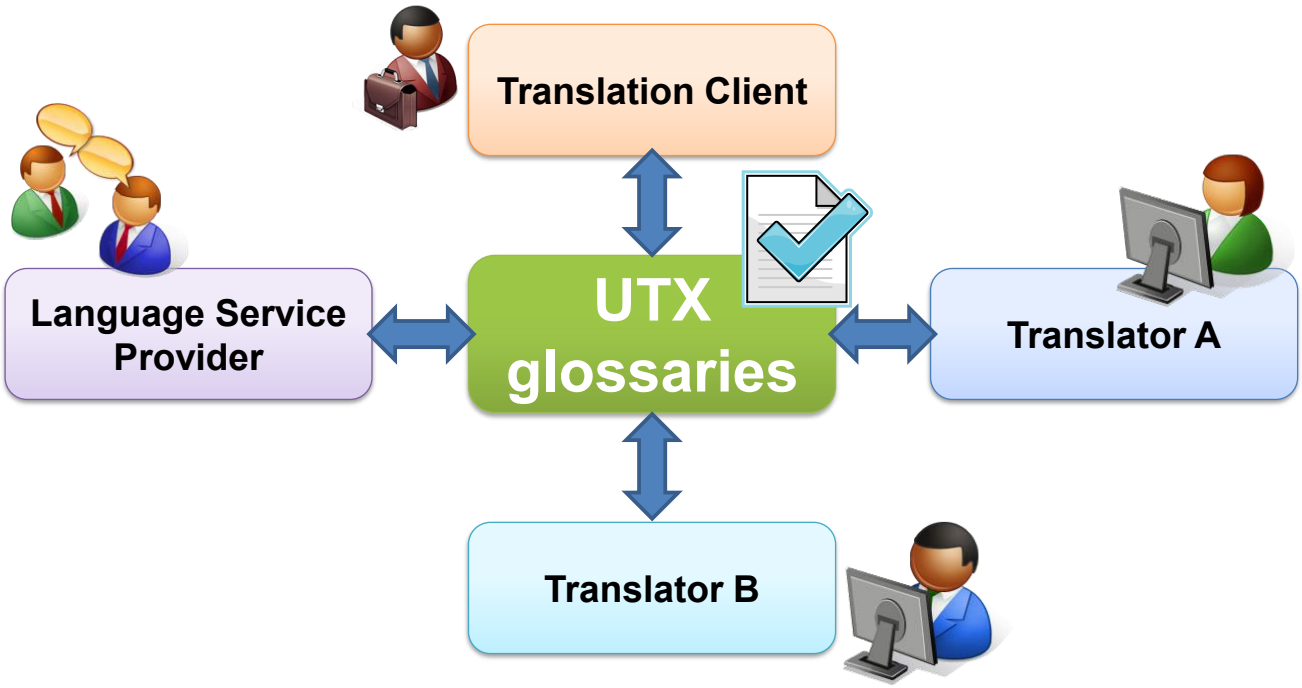
for terminology tools and MT

UTX is free to use

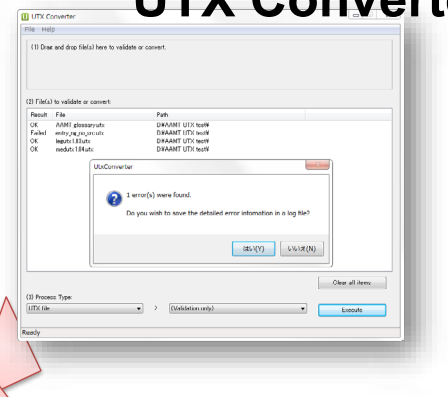
- UTX specification is free
- Many UTX glossaries are free
- UTX Converter is free (open source)



UTX facilitates sharing and reusing of glossaries



Conversion to/from UTX



UTX glossary sample

Manage essential glossary data in a standardized format

Information about the glossary (creation date, license, etc.)

#UTX 1.11; en-US/zh-CN; 2014-09-25; copyright: AAMT (2012); license: CC-by 3.0

#src	tgt	src:pos	term status
Asia-Pacific Association for Machine Translation	亚洲太平洋机器翻译协会	properNoun	approved
dictionary administrator	字典管理员	noun	approved
contributor	用语提交者	noun	provisional
domain	领域	noun	
glossary	词汇表	noun	
bidirectional	双向	adjective	approved
merge	合并	verb	approved

Source term (American English) | Target term (Chinese) | Part of speech | Term status

Term status provides **reliability**

JPO (Japan Patent Office) UTX glossary

- Created by JPO, converted by AAMT
- Available for free
- Japanese to English
- 130 thousands entries
- Only rare technical terms
  - User-defined terms not included in technical dictionaries shipped with the package

JPO glossary covers all IPC sections

IPC (International Patent Classification)

- A human necessities
- B performing operations; transporting
- C chemistry; metallurgy
- D textiles; paper
- E fixed constructions
- F mechanical engineering; lighting; heating; weapons; blasting
- G physics
- H electricity

Examples of entries

Domain (semantic feature)	Entries	Example	
Plants (common names, species, scientific names, etc.)	5498	白いぼキュウリ	white spine cucumber
		メラレウカ・アルテルフオリア	Melaleuca Alternifolia
		いらくさ科植物	urticaceous plant
Animals (common names, scientific names, etc.)	3025	ヤブカ	striped mosquito
		モンシロチョウ	Pieris rapae
		ユーグレナ	Euglena
People (personal names, titles, etc.)	1316	昌聰	Yoshiaki
		調香士	perfumer
		登壇者	presenter
Companies and organizations	7340	日本醸造協会	Brewing Society of Japan
		猟友会	hunters' association
		インド技術研究所	Indian Institute of Technology
Others	46975	Chemistry, medicine, machine, engineering, and other technical terms.	
		オキシジフタル酸二無水物	oxydiphthalic dianhydride

Patent documents characteristics

1. Extremely long sentences
2. Ambiguous sentence structure
3. Peculiar writing style
4. **Many technical terms (obfuscation)**

Terminology post-editing

What is “terminology post-editing”?

- post-editing method focused on terminology checking
- requires structured glossary data that has **strong correlation** with the source documents

Terminology post-editing: merits and limitations

■Merits

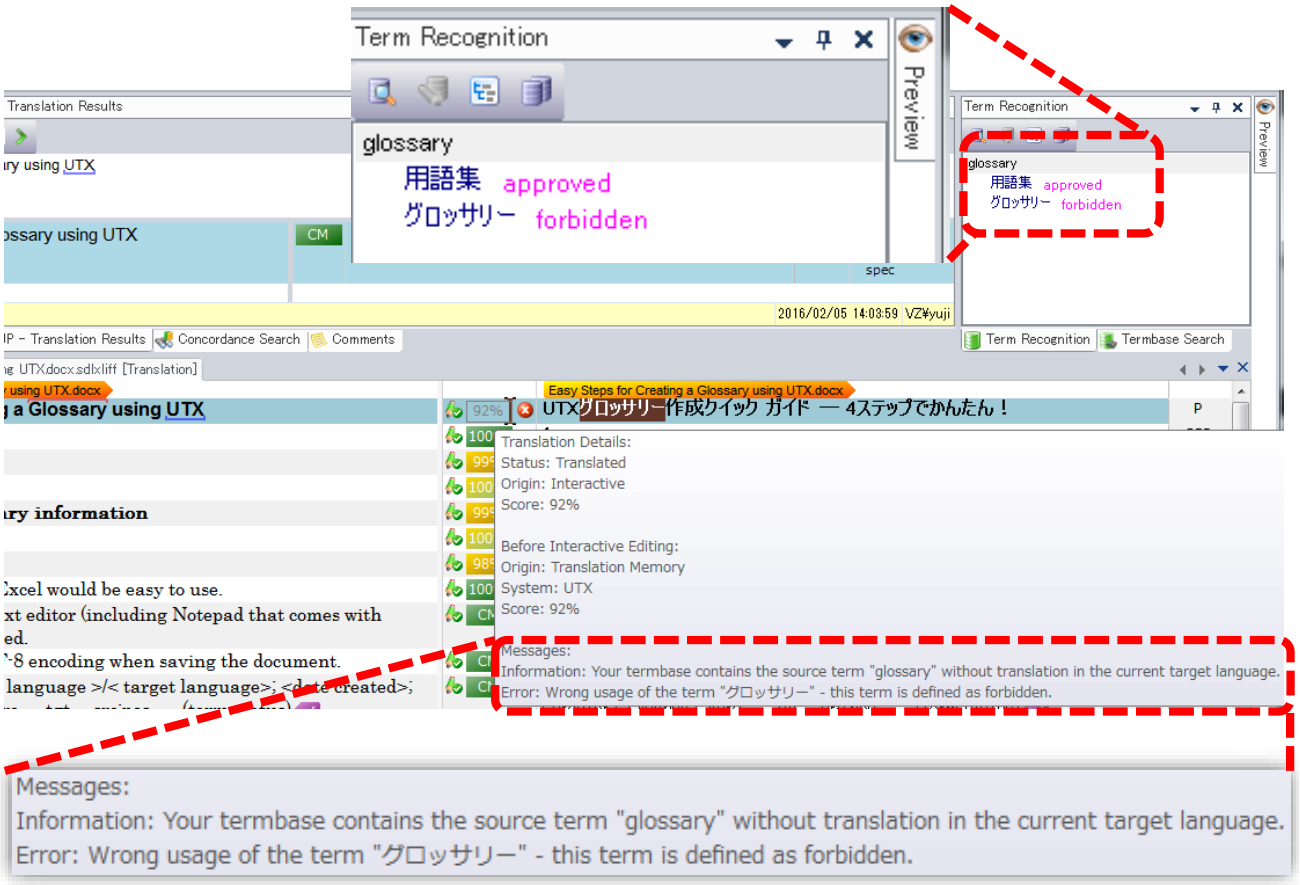
- Fully- or partially-automated check
- Check with no lingual knowledge

■Limitations

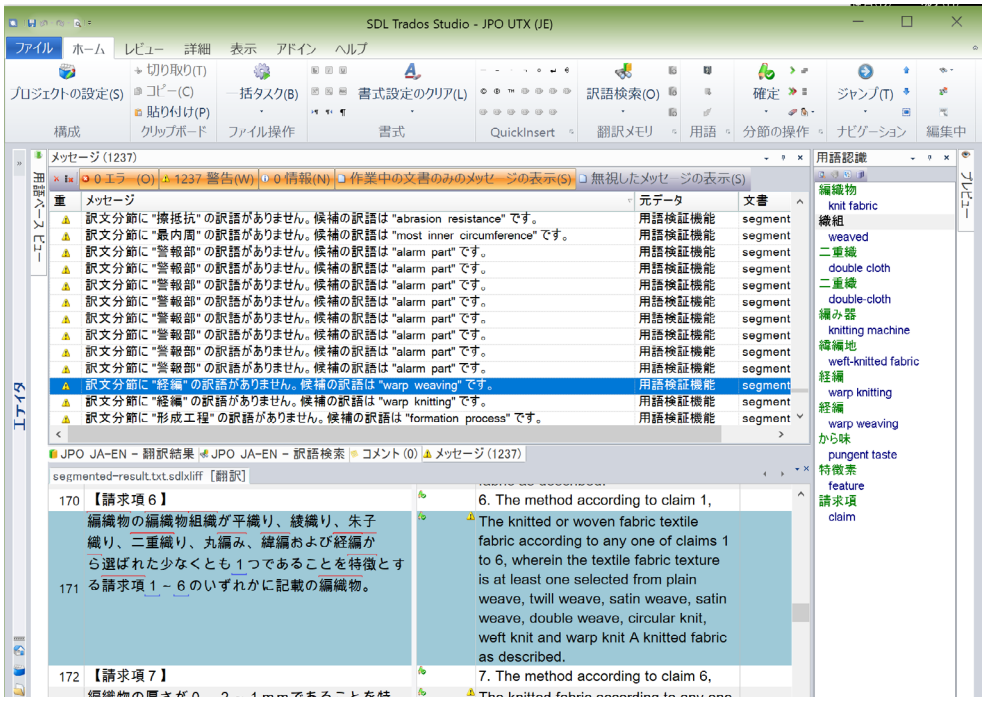
- Accuracy is insufficient
(requires other criteria for a full quality assessment)

Quick check or post-editing

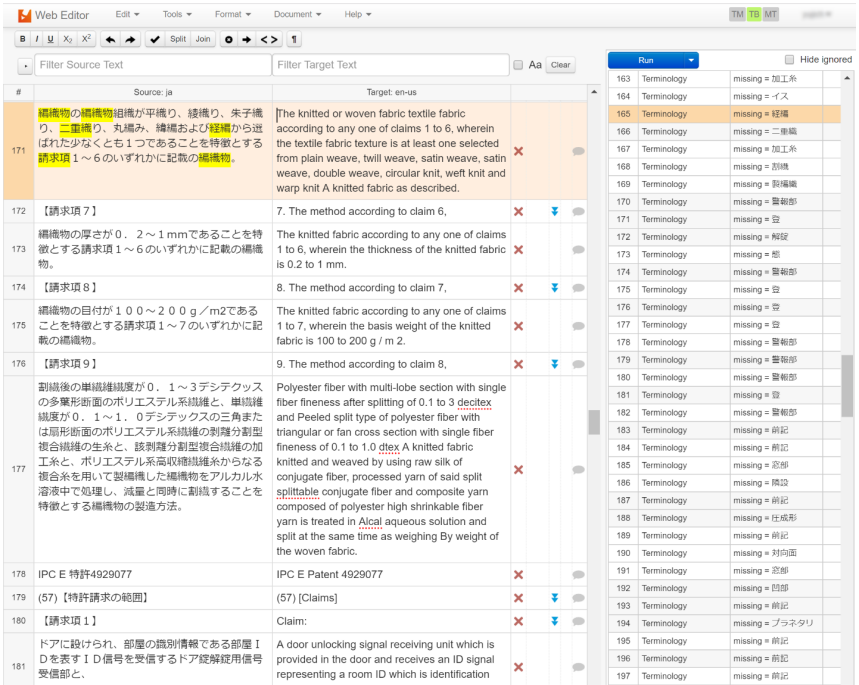
Terminology check in SDL Trados



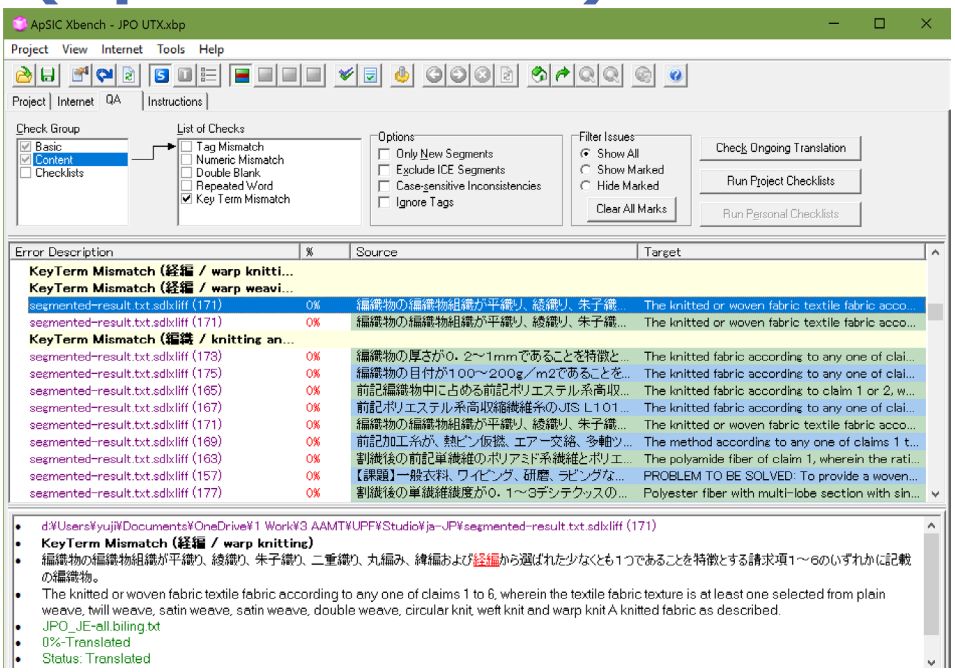
Patent NMT translation checked by UTX glossary data (SDL Trados)



Patent NMT translation checked by UTX glossary data (Memsource)



Patent NMT translation checked by UTX glossary data (ApSIC Xbench)



Result: potential term errors

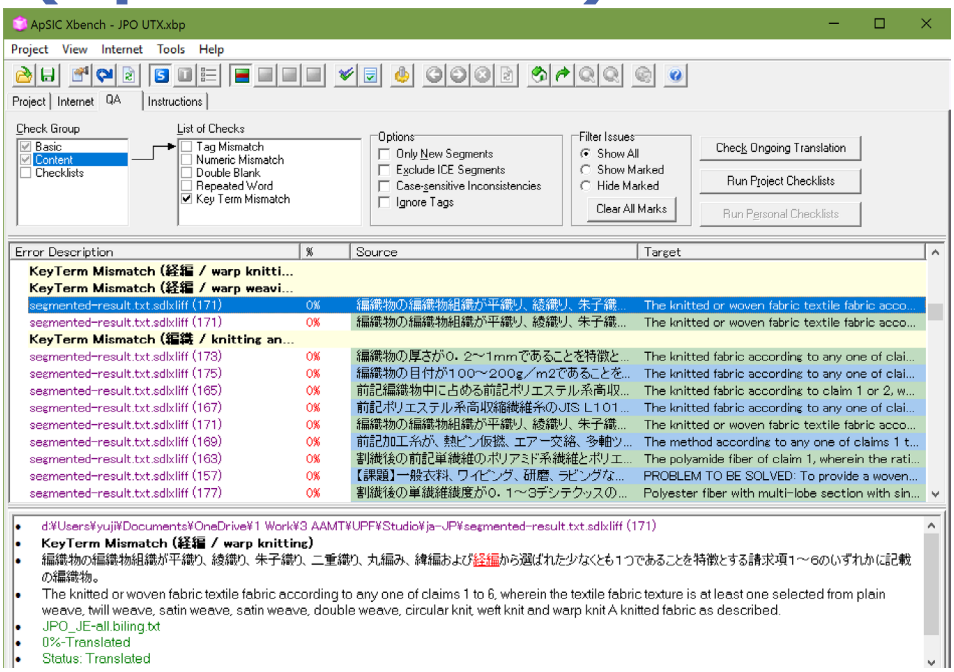
391 segments (sentences). More error detection is not necessarily better.

	Detected potential term errors
SDL Trados	1237
Memsource	372
ApSIC Xbench	603

■Examples of incorrectly translated terms:

- 請求項/claim
- デシテックス/decitex
- 経編/warp weaving
- センサシステム/sensor system

Patent NMT translation checked by UTX glossary data (ApSIC Xbench)



Result: potential term errors

391 segments (sentences). More error detection is not necessarily better.

	Detected potential term errors
SDL Trados	1237
Memsource	372
ApSIC Xbench	603

■Examples of incorrectly translated terms:

- 請求項/claim
- デシテックス/decitex
- 経編/warp weaving
- センサシステム/sensor system

Conclusion

1. NMT has many terminological flaws.
2. Glossary data and terminological check can find potential term errors.
3. To do so, you need a simple but structured glossary data format (such as UTX).
4. The UTX format was proved to be effective in finding potential errors.

Special thanks to Akimoto Kei (AAMT)

More info

■Visit http://www.aamt.info/english/utx/ for the specification and glossary data (free)

■Search for **"UTX glossary"**

■Contact at http://cosmoshouse.com/mail.htm

■We welcome your feedback!