Evaluations and Language resources Distribution Agency

9, rue des Cordelières
75 013 Paris
Tél. : 01 43 13 33 33
Fax : 01 43 13 33 30
Web : www.elda.org
Email : info@elda.org

Production Report of the TRAD Parallel Corpus Arabic-French – Translation of a subset of NIST 2008 Open Machine Translation Evaluation (LDC2010T21)

This corpus has been produced within the framework of the PEA-TRAD project (2011-2014)

Versions of the document :

Version	Date	Author	Status
1.0	02/07/2012	Djamel Mostefa, Priscille Schneller	Internal version (in French)
2.0	25/04/2017	Priscille Schneller	Final version (in English)

TABLE OF CONTENT

1 Introduction

This document is a report on the production process of the TRAD Parallel corpus Arabic-French (subset of LDC2010T21).

The source texts in Arabic were extracted from the NIST 2008 Open Machine Translation Evaluation corpus (LDC2010T21). A subset of about 20,000 words from the News domain has been translated into French within the PEA TRAD project (*TRADuction pour l'aide à l'analyse documentaire*, Translation as a Support for Document Analysis), supported by the French Ministry of Defence (DGA).

Translation into French has been conducted by ELDA and specific guidelines have been set up following ELDA's experience in similar projects.

This document gives an overview on the source texts, the translation / proofreading and validation process.

2 Description of the corpus

This parallel corpus contains 813 segments (translations units) from 74 documents.
The source file in Arabic contains 19,902 words and the reference translation in French contains 29,104 words.

The Arabic source texts were extracted from the MT08_Arabic-to-English subset of NIST 2008 Open Machine Translation (OpenMT) Evaluation sets (LDC2010T21), ISLRN: 415-534-082-867-8.

2.1 Data sources

The source texts are 'newswire' documents collected by LDC in July 2007 among news agencies and newspapers such as: Asharq Al-Awsat, Agence France-Presse, Al-Ahram, Al Hayat, Assabah, An Nahar, Al-Quds Al-Arabi, Xinhua News Agency.

The identifiers corresponding to the 74 documents extracted from LDC2010T21 are indicated below:

AAW_ARB_20070702.0001
AAW_ARB_20070703.0024
AAW_ARB_20070704.0025
AAW_ARB_20070707.0027
AAW_ARB_20070710.0019
AAW_ARB_20070711.0107
AAW_ARB_20070715.0042
AAW_ARB_20070720.0044
AAW_ARB_20070722.0070
AAW_ARB_20070723.0003
AAW_ARB_20070723.0048
AAW_ARB_20070727.0002
AAW_ARB_20070728.0039
AAW_ARB_20070730.0057
AAW_ARB_20070731.0008
AFP_ARB_20070708.0015
AFP_ARB_20070710.0147
AFP_ARB_20070714.0026
AFP_ARB_20070716.0037
AFP_ARB_20070718.0054
AFP_ARB_20070718.0137
AFP_ARB_20070719.0103
AFP_ARB_20070719.0133
AFP_ARB_20070722.0026
AFP_ARB_20070722.0102
AFP_ARB_20070723.0091
AFP_ARB_20070726.0119
AFP_ARB_20070730.0074
AFP_ARB_20070731.0138
AHR_ARB_20070702.0037
AHR_ARB_20070716.0038
AHR_ARB_20070716.0098
AHR_ARB_20070726.0093
AHR_ARB_20070728.0029
AHR_ARB_20070730.0042
AHR_ARB_20070730.0063
AHR_ARB_20070731.0033
ASB_ARB_20070730.0012
HYT_ARB_20070702.0060
HYT_ARB_20070704.0004
HYT_ARB_20070704.0031
HYT_ARB_20070706.0035
HYT_ARB_20070711.0064
HYT_ARB_20070714.0015
HYT_ARB_20070715.0048
HYT_ARB_20070719.0012
HYT_ARB_20070720.0014
HYT_ARB_20070722.0070
HYT_ARB_20070726.0057
HYT_ARB_20070730.0022
NHR_ARB_20070711.0073
NHR_ARB_20070712.0035
NHR_ARB_20070716.0051
NHR_ARB_20070720.0102
NHR_ARB_20070730.0059
NHR_ARB_20070731.0060
QDS_ARB_20070702.0030
QDS_ARB_20070710.0011
QDS_ARB_20070714.0038
QDS_ARB_20070724.0030
QDS_ARB_20070728.0048
QDS_ARB_20070730.0023
XIN_ARB_20070702.0073
XIN_ARB_20070703.0198
XIN_ARB_20070704.0086
XIN_ARB_20070708.0053
XIN_ARB_20070712.0073
XIN_ARB_20070714.0122
XIN_ARB_20070718.0039
XIN_ARB_20070719.0204
XIN_ARB_20070722.0127
XIN_ARB_20070726.0145
XIN_ARB_20070728.0034
XIN_ARB_20070731.0042

2.2 Data format

The corpus contains 2 XML files:

- nist08nw_ar_src.xml
- nist08nw_fr_ref.xml

Files are named according to the following rules:

<corpus_id>_<lang>_<id>.xml

where:

- <corpus_id> is the name of the corpus
- <lang> = ar | zh | fr | en is the document's language
- <id >= src | ref is the identifier of the document ('src' as 'source' and 'ref' as 'reference translation')

The files are unicode-encoded XML following the DTD provided in the Appendix p.12.

3 The translation process

The objective of the current work is to produce high-quality bilingual data being the result of the work of translation professionals. The produced corpora are used to train or evaluate machine translation systems.

The composition of the translation team and the specification of translation guidelines are defined by ELDA at the beginning of the project to regulate the translators' work.

3.1 The translation team

A single translation team should translate all of the source language data. This team might be composed of:

- one or several bilingual translator(s), native speaker(s) of the target language of the data.
- a bilingual target language native speaker who proofreads and edits the output of the translators. He is also in charge of the homogenisation of the whole corpus, especially regarding terminology, if required. Notice that the translations must be systematically finalised and checked by a target native speaker.

Each file is encoded in XML and identified with a series of information that does not need to be translated: document version, encoding, DTD, source of the data as well as language and system ID.

Each document contains one or several main fields with text to translate: these are contained within XML tags and are referred to as segments (e.g., <seg id="1" >).

The file is to be rendered as XML format, UTF-8 encoded, so as to preserve the original structure.

3.2 Translation quality

Translation agencies follow their best practices to produce translations. While we trust that each translation agency has its own mechanisms of quality control, we have specific guidelines so that all translations share a common ground. These are:

1. The target translation must be faithful to the original source text in terms of meaning and style. When the source text is a press dispatch, the translation should be written in a journalistic style, thus respecting the document style. When the source text is taken from blogs or mailing lists, unusual syntactic structures may be used. The translation should reflect this phenomenon. The translation should mirror the original meaning as much as possible without compromising grammaticality, fluency and naturalness.

2. The tone and register of the language should be respected. For instance, if the text shows an angry or uneasy speaker in the source language, this state of mind should be also expressed in the target language, conveying the same tone.

3. The same applies for the general "politeness" and "formality" register of the source text. Both translators and proofreaders should bear in mind the "politeness" standards of the target language, such as the use of pronouns "vous" and "tu" in French when translating from English.

4. The translation should be as factual as possible, trying to keep the exact information conveyed by the source text, without changing the meaning or without adding/removing information. For example, if the original text uses "Obama" to refer to the U.S.A. President, the translation should not be rendered as "President Obama", "Mister Obama", etc.

5. No bracketed words, phrases or other annotation should be added to the translation as an explanation or aid to understanding.

6. The translation should entail the same cultural assumptions as the original text, and no implicit reference should be made explicit by the translator.

7. The order of consecutive segments must not be altered, not even for stylistic reasons, i.e. the contents of segments N and N+1 must not be swapped in the translation.

8. Should a sentence to be translated prove to be incomplete (due to formatting problems, etc.) and pose problems for its translation, it should be reported to ELDA immediately. Then it will be decided whether such sentence must be corrected or deleted.

9. Capitalisation and punctuation are language dependent. This means that translators should follow the standards from the target language and apply their rules even if these may not coincide with those of the source document. This may imply, for instance, changing or adding either upper/lower case letters or punctuation marks.

10. Regarding neologisms and unknown words: if it is possible to understand the intention/gist of the source text, then the translation should be either the correct form of the word (for unknown words) or a new word corresponding to the source derivation (for neologisms). If the translator has no preexisting knowledge on how to translate a word, (s)he is expected to consult standard sources, such as dictionaries, translation forums, etc.

11. Regarding proper names, whenever possible, these should be translated following conventional practices in the target language. However, it may be necessary not to follows instructions, as for instance concerning the order of first names and last names, which must be translated according to source language’s standards. As with neologisms, when lacking knowledge on the word to translate, translators are expected to consult standard resources.

12. Regarding titles (movies titles, series, books), their translation must follow the standards of the works’ translations in the target language. If these standards don’t exist, titles must be left untranslated in the form that can be found in source language. For instance, “Voice of America” should be translated into French by “La voix de l’Amérique”, which is the standard translation.

13. Translators are requested to list all cases (proper names, neologisms, etc., as indicated in points 10 and 12) where such standard resources have been used. This will help to assure the consistency of the translation of the completed document.

14. The format of entities like dates and numbers in general must remain the same in the translated document. For instance, when translating from English into French, “January 27, 1999” must be translated into “27 janvier 1999”, while “January twenty-seventh nineteen ninety nine” must be translated into “vingt-sept janvier mille neuf cent quatre-vingt-dix-neuf”

15. Idioms and colloquial expressions are particularly hard to translate. If a similar expression exists in the target language, it should be used. However, if there is no direct translation into the target language, try to preserve the meaning of the source-language expression but convey it in as natural and fluent a target-language expression as possible.

16. If spelling errors are to be found in the source text, they must be ignored and must not be reproduced in the translation.

17. The normalisation and revision of the whole corpus will be done in terms of terminology used, as well as orthographic consistency, style and register. For consistency purposes, the proofreading of the full corpus will be done by a target language native speaker. A report should be produced by the proofreader to make explicit the improvements introduced during the process.

4 Validation

4.1 Validation protocol

To assure the quality of the translations, ELDA enforced the following policies:

1. Hire fluent bilingual experts to control the translation quality level.

2. For each delivery of translated texts, a subset of 5% of the documents is randomly selected and the selected sample translation is graded.

3. To ensure consistency from one review to another, a system has been adopted for grading translations:

Error	Penalty
Syntactic	3 points
Lexical	3 points
Poor usage of target language	1 point
Capitalisation / Spelling error	1 point
Punctuation	0.5 points (max 10 points)

- **Syntactic errors** are those found in grammatical categories. These comprise errors such as problems with verb tense, coreference and inflection. Furthermore, syntactic errors are also those where there has been a misinterpretation of the grammatical relationships between the words of the original text. Examples of syntactic errors are, for instance, translating an object as a subject, making an adjective modify a verb, attaching a relative pronoun or prepositional phrase to the wrong noun.
- **Lexical errors** comprise omitted words or wrong choice of lexical item (word), due to misinterpretation or mistranslation.
- **Poor usage** of target language means awkward, unidiomatic usage of the target language and failure to use commonly recognised titles and terms.
- **Capitalisation errors** refer to the initial character of a sentence, as well as any words which do not respect the capitalisation conventions in the target language. For instance, proper names should start with upper-case in certain languages and this should be taken into account when translating into such target languages.
- **Punctuation errors:** punctuation should also follow the standards/conventions of the target language, even if the source language is not correctly punctuated.

4. Over a defined level of penalty points, the translation is rejected and the whole delivery sent back to the translation team for improvement. Once the translation is corrected, a new sample is extracted and validated under the same protocol.

The total of penalty points in the sample must be lower to the defined threshold for the translation to be accepted. The hierarchy of errors shows errors’ gravity and applies penalties commonly used in the domain and in similar machine translation projects such as CESTA, TC-STAR, QUAERO, MEDAR.

The acceptance threshold here is 1 point for 100 words.

4.2 Validation results

The sample extracted for validation contains 1500 words.
Our acceptance threshold here is 15 points.
The table below shows the validation results:

Error	Number of errors	Penalty points
Syntactic	1	3
Lexical	0	0
Poor usage of target language	4	4
Capitalisation / Spelling error	0	0
Punctuation	4	2
TOTAL		9

The validation results above are related to the last version of the translation.

4.3 Revision

When the translation is rejected, the translation team is asked to revise it in its entirety. Comments from the validation team must be taken into account and the improved translation completed within an agreed time-frame.

Before final delivery, validation tools are used to check format and detect remaining spelling errors on the whole translation.

Appendix: XML format DTD

```
<!--
 History:
 version 1.4: adaptation to PEA TRAD languages
 version 1.3: added English and Urdu as srclang values
    added Chinese as trglang values
    added poster element
    made the sysid attribute required

 version 1.2: converted from a SGML DTD to an XML DTD,
    added new attribute genre to the doc element

 version 1.1:  the NIST DTD for evaluation of language translation
    adapted from LDC DTD for Multiple-Translation Chinese Corpus,
    version 1.0
-->

<!ENTITY lt     "&#38;#60;">
<!ENTITY gt     "&#62;">
<!ENTITY amp    "&#38;#38;">
<!ENTITY apos   "&#39;">
<!ENTITY quot   "&#34;">

<!ELEMENT mteval (srcset | refset+ | tstset+)>
<!ELEMENT srcset (doc+)>
<!ATTLIST srcset setid CDATA #REQUIRED>
<!ATTLIST srcset srclang (Arabic | Chinese | English | French | Pashto)
#REQUIRED>
<!ATTLIST srcset trglang (English | French |Pashto) #REQUIRED>

<!ELEMENT refset (doc+)>
<!ATTLIST refset setid CDATA #REQUIRED>
<!ATTLIST refset srclang (Arabic | Chinese | English | French | Pashto)
#REQUIRED>
<!ATTLIST refset trglang (English | French |Pashto) #REQUIRED>

<!ELEMENT tstset (doc+)>
<!ATTLIST tstset setid CDATA #REQUIRED>
<!ATTLIST tstset srclang (Arabic | Chinese | English | French | Pashto)
#REQUIRED>
<!ATTLIST tstset trglang (English | French |Pashto) #REQUIRED>
<!ELEMENT doc (hl | p | poster | seg)*>
<!ATTLIST doc docid CDATA #REQUIRED>
<!ATTLIST doc genre (news | speech | editorial | text) #REQUIRED>
<!ATTLIST doc sysid CDATA #REQUIRED>

<!ELEMENT hl (seg*)>

<!ELEMENT p (seg*)>

<!ELEMENT poster (seg*)>

<!ELEMENT seg (#PCDATA)>
<!ATTLIST seg id CDATA #REQUIRED>
```