The A to Z of Manually Translated Parallel Corpora for the Evaluation of Machine Translation: Establishing Protocols through Experience. Session 5 - Assessing Quality in MT

Victoria Arranz, Olivier Hamon, Karim Boudahmane, Martine Garnier-Rizet

TRALOGY

The A to Z of Manually Translated Parallel Corpora for the Evaluation of Machine Translation: Establishing Protocols through Experience

Victoria Arranz
ELDA
{arranz,hamon}@elda.org

Olivier Hamon
ELDA

Karim Boudahmane
DGA
karim.boudahmane@dga.defense.gouv.fr

Martine Garnier-Rizet
IMMI
garnier@immi-labs.org

TRALOGY II - Session 5
Date d'intervention : 18/01/2013

The evaluation of Machine Translation (MT) systems often requires the use of reference translations in order to compare the MT system output against those translations. This comparison allows measuring the quality of the automatic translations. Such reference translations are generally produced by professional translators who follow a series of well-defined protocols and guidelines defined for the specific evaluation data (or corpora) to be produced. The current paper aims at describing in detail how the entire evaluation-data production procedure is carried out. This will imply going through the different stages and needs of the project, the constitution of guidelines, the setting up of protocols and teams, etc. All of it is done bearing in mind that one such project is a dynamic entity and that it evolves as requirements change, feedback from teams is received and as we learn to optimise it through experience.

lien video : http://webcast.in2p3.fr/videos-unsupervised_estimation_of_mt_quality

1. Introduction

The evaluation of Machine Translation (MT) systems is often linked to an automatic comparison with one or several translations produced by professionals. These are called reference translations. Current best-known automatic evaluation measures, such as BLEU [Papineni, K. et al. (2001)], NIST [Doddington, G. (2002)] or METEOR [Banerjee, S. and Lavie, A. (2005)], together with most of the current measures [Callison-Burch, C. et al. (2010)], provide a comparison between reference translations and machine translations.

Corpus production may be done either with source texts translated by professionals, or by collecting documents that are already translated [Koehn, P. (2005)] and that are found, for instance via *Internet* and that are aligned at sentence level. Both methods have their own advantages and drawbacks. The former is generally "accused" of being more expensive, however, it allows to obtain more reliable translations given that the translator is well acquainted with the translation context and domain. Besides quality, a further issue of concern when collecting bilingual data, in particular from the *Internet*, is that of obtaining comparable rather than parallel corpora.

Bearing this in mind, evaluators require quality human translations that respect strict constraints regarding both the context of the source documents and the translation task.

This past decade, ELDA has been producing translated parallel corpora both for the development of Machine Translation (MT) systems and as reference for their evaluation. All throughout this time, our goal has been to support the production of language resources (LRs) for the design, development and evaluation of MT systems. Taking as starting point European projects like TC-STAR [Mostefa, D. et al. (2007)] and MEDAR [Hamon, O. and Choukri, K. (2011)], together with the French CESTA project [Hamon, O.et al. (2007)], the objective has always been to produce high-quality bilingual data[1], being these the result of the work of translation professionals (translators). Our production work has not only been the outcome of project requirements, but also that of a thorough analysis and consideration of both data users' needs and translators' feedback.

When producing resources, the quality of the result is essential: procedures and rules must be formalised through guidelines and documentation.

This production of parallel corpora is neither new nor restricted to ELDA. A number of national and international projects use parallel corpora for MT system evaluation and a number of companies are using translation daily in order to produce their data. Most of the translations are the result of work following guidelines, such as the data produced within the DARPA GALE programme[2] or within the NIST Open MT Series[3].

However, these production procedures are never static statements or definitions. Feedback from all teams as well as lessons learnt during the process are crucial for success. Thus, a key point throughout all these years of data production has been the regular updating of the procedures and their associated guidelines to tackle all predictable data production phenomena.

The definition of guidelines is done at different levels. The first level regards the *recruitment of the translation team(s)*: professionals are selected following several criteria and subject to meeting quality expectations defined a priori. The second level concerns the *translation guidelines* defined so as to help translators translate a specific language direction, for a particular domain and in a certain context. Finally, there are the *validation guidelines*, which establish the criteria and measures to use in order to check as objectively as possible the quality of the trans-

(1) The evaluation data produced and described in this document will be also referred to as "corpus/corpora" and "reference(s)" throughout the document.
(2) http://projects.ldc.upenn.edu/gale/Translation/
(3) http://www.itl.nist.gov/iad/mig//tests/mt/

lation produced. At ELDA, those guidelines have been regularly adapted to meet different needs, according to the domain of the data (news, politics, medicine, patents, etc.), their nature (text, transcription), their structure and format (plain text, DTD-structured XML data, etc.) or the translation direction (e.g. English- to-Spanish, Arabic-to-English, German-to-French, Chinese-to-English, among others).

These past few years, this guideline scenario has been faced with one further step: the handling of issues raised by the segmentation of the source audio data. This has required that a new step be put into place for data production and, as a consequence, the necessary guidelines be also defined for such a task, i.e. audio data *re-segmentation guidelines*.

This A to Z for the production of translated parallel corpora does not pretend to establish a completely new procedure, but rather to define a clear and up-to-date one, matured with ELDA's experience and expertise throughout the years. As mentioned earlier, this work is based on earlier definitions of procedures and documents, mostly derived from the TC-STAR[4] and CESTA[5] experiences, as well as from the rich work carried out within GALE. Such experiences have been tuned and refined for the evaluation data production carried out for the DGA[6] (Direction Générale de l'Armement) and the IMMI[7] (Institute for Multilingual and Multimedia Information) these past few years in the framework of the Quaero project[8].

This paper draws a detailed how-to of the production of translations for the creation of parallel corpora, explaining the full procedure to manage such a production project. This involves taking into account how teams are constituted, how guidelines are defined and aims at sharing the experience and lessons learnt throughout these years on several aspects such as: problems encountered when handling transcription data, establishing quality criteria, among others. In this regard, the paper starts by describing the pre- translation steps, then moves onto translation and every aspect concerned, closing up the procedure with the quality control implemented throughout the procedure. The final sections of the article give an overview of the problems encountered when handling spontaneous speech data together with the lessons learnt and adopted solutions.

2. Corpus Production Procedure

When talking about the production of translation references for technology evaluation, it may sound like all there is to it is the correct translation of source texts. However, before going into the translation part of the data production work, some tasks must be taken care of carefully. These are essential for the success of the whole project. Such tasks concern the points described in the subsections below.

2.1 Definition of the Data and its Features

When a new production project is about to start, a number of key characteristics need to be defined: data domain and nature; languages involved and translation direction; data size and delivery-timing demands. All of these are clearly established as they will have an impact on the following steps. In the context of the current work, both text (e.g. journalistic, patents) and audio data (e.g. radio and TV, debates, parliamentary speeches) types and domains have been handled. The languages treated comprise Arabic, Chinese, English, French and German, in a number of direction combinations. Regarding data size, source corpus size ranges between 22,000 and 27,000 words, the former concerning the audio data and the latter the text one. The reason for these sizes is that evaluation corpora do not need to be as large as training

(4) http://www.tcstar.org
(5) http://www.technolangue.net/imprimer.php3?id_article=199
(6) http://www.defense.gouv.fr/dga
(7) http://www.immi-labs.org
(8) http://www.quaero.org/modules/movie/scenes/home/

data (which may contain thousands and even millions of word, if available), but rather focus on having a wider variety of references (a minimum of 2 but with a preference for more). Moreover, evaluation data are required to be of a very high-quality, which means the procedure to produce them is longer and more costly, which for budget reasons, imposes some conditions on their sizes. Last but not least, and from a realistic point of view, evaluation data are generally produced for a particular technology evaluation campaign, which sets up very strict deadlines in terms of production and thus, does not allow the creation of very large corpora.

It may be worth mentioning that a few years back, ELDA carried out some work on the production of evaluation data based on the paraphrasing approach. This consisted in creating 16 different versions of each sentence in the source corpus, providing different ways to say or write a sentence. The source sentences were paraphrased themselves, so as to help in the production. The richness achieved in terms of evaluation data was very interesting, which was the aim for that particular evaluation, but the challenge behind it was often exhausting for the translators. The source data sometimes contained rather short sentences that the translators found very difficult to translate in 16 different ways, even with the help of the source paraphrasing. One example of such data is as follows:

English source:

I'd like to go here.

French target translations and paraphrases:

01\Je veux aller ici.
02\J'aimerais aller ici.
03\C'est ici que je veux aller.
04\J'aimerais aller à cet endroit.
05\C'est à cet endroit que je veux aller.
06\Je veux aller à cet endroit.
07\Voici l'endroit où j'aimerais aller.
08\J'aimerais aller là.
09\C'est là que je veux aller.
10\Voilà l'endroit où je veux aller.
11\Je voudrais aller ici.
12\C'est là que je voudrais aller.
13\Voici l'endroit où je veux aller.
14\J'aimerais y aller.
15\Je souhaiterais aller à cet endroit.
16\Je voudrais aller à cet endroit.

3. Recruitment of a Translation Team according to the Specific Needs

The recruitment of the translation experts follows the criteria listed in the previous section, the biggest constraints being the language and domain expertise of the candidates. Needless to say that the implications of translating, for instance, Chinese pharmacological patents into English have nothing to do with working on German radio broadcast data to be translated into French. These specificities define the profile of the experts to be taken in the project, who will be either already known language specialists (with whom we have already worked) or who will be duly tested for the task. Every time a new language specialist (this applies to all translators, proofreaders and validators) is considered, a test is carried out to evaluate his/her competence and thus make sure (s)he is appropriate for the task.

4. Recruitment of a Validation Team according to the Specific Needs

Bearing also the above-example in mind, the quality control in the Chinese patent translation counts on the expertise of a translator, who is both specialised in that language direction and holds a degree in the pharmacological field. Validating such type of documents and terminology is an extremely delicate matter, not only because of the complex and specific terminology, but also due to the particular and complex syntax behind these patents. As mentioned in the earlier section, validators are either part of our already-established group of language experts or they will need to go through a test (specific for their task) to prove they are suitable.

5. Formatting of the Source Data

Depending on the type of data to handle, the formatting to carry out on the source data may increase in complexity. A number of input formats are handled, including audio formats such as TRS[9], and we generally aim at obtaining "simple" XML source files that can be easily shared and used by a variety of software and users. Let us not forget that translators use translation aids and expect/prefer input files as easy to handle as possible. Furthermore, when working on the translation of audio data, this may also mean the filtering or removing of certain or all speech phenomena, transcription tool tags, etc. that do not need to be translated while keeping information (such as segment timings) that will be of use for the Speech Language Technology (SLT) tools to be evaluated.

6. Re-segmentation of the Source Data

The re-segmentation of the source data may also be an additional step when creating references for audio transcriptions. The goal of the re-segmentation procedure is to prepare transcription data so as to be translated into another language. Already transcribed data may need to be re-segmented in order to obtain well-formed and self- contained sentences. This is particularly important when translation is to be done between languages with a very different syntactic structure (such as French and German, where the latter, as opposed to the former, places the verb at the end of the sentence), and where fragmented sentences pose serious problems for translation and their later alignment. Thus, the task aims at providing translators with complete (or as complete as possible) sentences, by either regrouping broken segments or splitting several sentences which have been inserted into the same segment.

7. Planning and Cost Estimate of the Production

The timing-delivery and budget constraints have already been mentioned earlier on as having an impact on the constitution of the project. As it has been explained, the production requirements, in general, set up the lines to follow along the project. However, there are two further points to be taken into account in terms of timing and finances: the planning and the cost estimates of the production, both of them closely related to the quality expectations and management required to achieve them. Besides the translation cost (which may vary according to various parameters), the supervision or management required to lead a project to success is an expense very often neglected when designing a project. However, the efforts required are sometimes large, increasing, for instance, with the number of necessary validation stages. The higher the quality expected, the more time and effort consuming the full production will be.

(9) TRS is the extension for the files created using the speech transcription tool Transcriber: http://trans.sourceforge.net/en/presentation.php.

Generally speaking, we can say that an average of 50/60 working days are needed for the full translation procedure of an about 22,000 word source corpus. This may need to be readjusted for a number of reasons, such as validations returning low scores and thus needing to do further revisions and new validations. Further details on this will be provided in the "Quality Control" section.

8. What Next?

As it has been seen, the before-explained tasks are of a different nature, but all interrelated and having an impact on the following pre-translation tasks:

> Definition of re-segmentation guidelines. Definition of translation guidelines. Definition of validation guidelines.

Following all these preliminary tasks, the documents are sent to the translators. Translation teams are always encouraged to approach the project leader whenever questions or unclear aspects arise in the guidelines. Indeed, producing translations for the development or evaluation of NLP tools is not a trivial task for translation professionals who are generally not used to produce such data. This is even worse in the case of validation as translators are sometimes reluctant to check the work of other translators when they are meant to pinpoint errors according to our guidelines.

Once the translations are produced and validated, data are packaged so as to be used by the systems. Non validated (rejected) data are sent back to the translation team for proofreading. In some cases, rejected data can be returned with corrections and/or comments from the validator so as to guide the proofreader in the types of issues to be particularly aware of.

The general workflow for the data production described in this paper looks as illustrated in Figure 1. The external parts of the workflow comprise the additional re-segmentation task that can be required for the treatment of the audio data together with its guidelines (in red) and the final usage of the data produced: the MT Evaluation (in green-bleu).

The coming sections will give an overview of the criteria and parameters which are defined within both translation and validation guidelines.

9. Translation Protocol

9.1 Setting up the Translation Team

As mentioned in the introduction, the translation team is carefully recruited following a number of criteria defined a priory:

- A single translation team is used to translate all of the source language data and produce one single reference. This team is composed of:

1. One bilingual translator, native speaker of the target language of the data.

2. A target native speaker bilingual who proofreads and edits the output of the translators. He may be also in charge of the homogenisation of the whole corpus, especially regarding the vocabulary, if required (see point 3).

3. Should data size and time constraints impose so, several bilingual translators may be allowed to work on the same source data, knowing that the homogenisation stage will be then required.

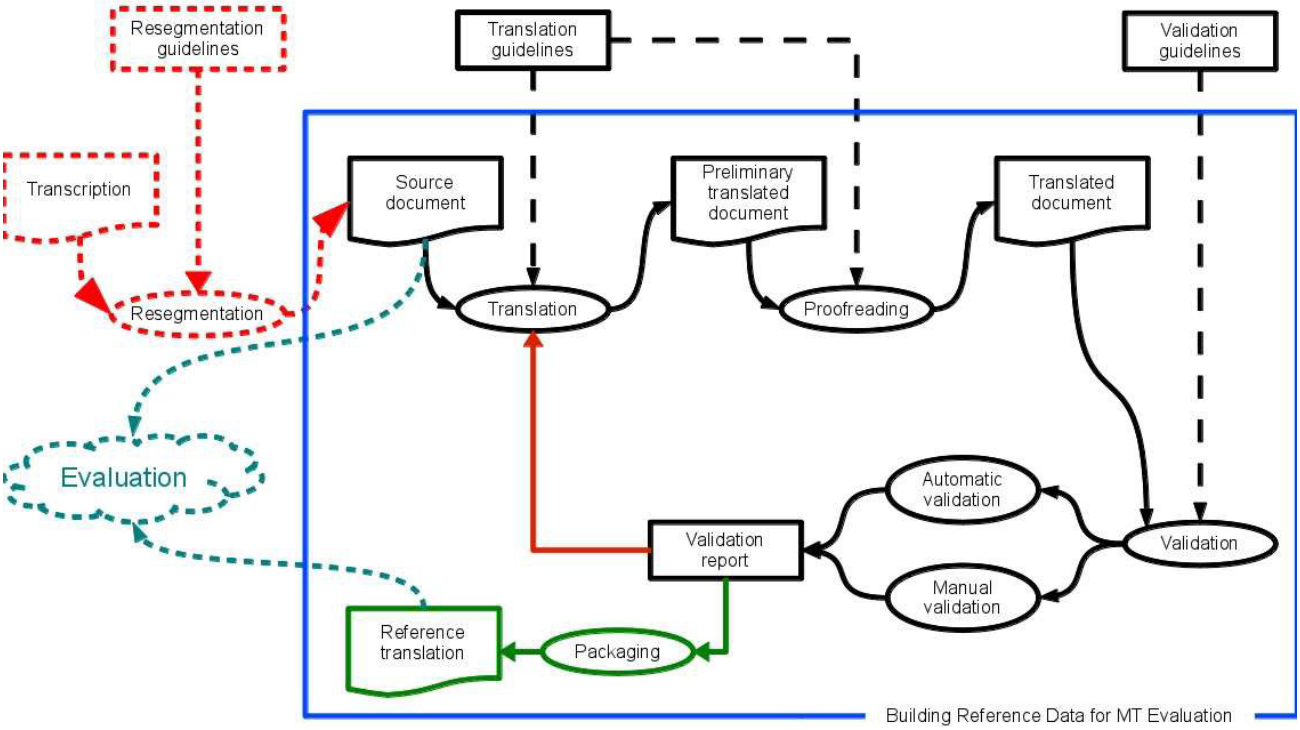

Figure 1: Reference Data Production Workflow

- It should be further noticed that:

Translations must be systematically finalised and checked by one target native speaker.

The translation team should not change during the course of translation.

The translation team must be fully documented, providing detailed information on each member's name (or pseudonym), native language, second language(s), experience, etc. This information may be anonymised when working with translation agencies as, in practice, most agencies object to disclosing certain details about their translators, in order to avoid any potential direct communication between their translators and their customer.

Translators are also asked to report on any additional quality control procedures they implement as well as on any other relevant parameters or factors that affect the translation. For instance, when working on audio data translation, besides incurring into discussions on how to handle certain sentences they will also provide us with specific problematic cases or situations that may bring up some questioning. This will be their way of justifying their choices.

Last but not least, teams become "established experts" over the years for specific languages, domains and data types. With regard to the transcription data, this is highly valuable as it is very complex data to handle and translators are duly trained to work on it. In addition to the translation guidelines provided, feedback is given on a regular basis to coach the work to be done. Moreover, in a number of occasions, translators have refused to translate such data explaining that they would not work on "incorrect sources" (audio data are full of hesitations, partially pronounced words, and other such phenomena) or claiming that one such translation was rather the work of interpreters.

9.2 Source Data Definition

The definition of the source data to be translated is an important section of the guidelines. It explains exactly what the source looks like and how the target translation should be also in terms of format, structure and encoding. Generally, source data are monolingual texts coming from different domains, such as newspapers, news web sites, commercial and scientific patents, etc. Data can also be of any nature that may require an adaptation of the guidelines as, for instance, transcriptions derived from audio data. This is all analysed in advance so as to adopt the relevant guidelines (different documents have already been produced) and adapt them accordingly in order to clearly define the work of the translation team and thus avoid ambiguous points.

In general, each file is encoded in XML, identified with a docid attribute and a language code, although that may differ according to customer's needs. If some information does not need to be translated (e.g., document version, encoding or DTD used), this is indicated. Each paragraph is tagged and may contain several sentences. Within the paragraph tags, one sentence must correspond to one separate line.

Once translated, the file is to be rendered as XML format, UTF-8 encoded, so as to preserve the original structure. If further treatment is needed towards the preparation of the aligned evaluation references, this is part of ELDA's post-translation data preparation step. Translators are just asked to preserve the source format so as to simplify their task.

9.3 Translation Criteria

All throughout these years, ELDA has developed and improved translation guidelines to control the quality of the documents returned by translators. Even if best practices are used by translators and we trust that each translation agency has its own mechanisms of quality control, specific points still need to be cleared in advance. The use of guidelines helps everyone involved in the translation procedure to share a common ground, being based on procedural principles and criteria. Nevertheless, it should be pointed out that despite the use of guidelines, obstacles still arise and translators are faced with ambiguities, misunderstandings and disagreements between the teams that we often need to resolve on a case-per-case basis.

Below follow some of the translation criteria defined in the guidelines so as to illustrate the kind of information the translators base their work on. For example, some of them focus on style, tone, register:

- 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. The translation should mirror the original meaning as much as possible without sacrifying grammaticality, fluency and naturalness.
- 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.

Others on translation fidelity and sentence order:

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

- 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.

Specific content issues are also addressed, where instructions are offered for those cases that may be particularly problematic and ambiguous:

- Regarding the translation of titles (for books, TV series, films, etc.) translators are expected to use standardised translations. If such standardised versions do not exist, titles should be left untranslated, as in their source language.

Normalisation of the output is also instructed for the translation team:

- 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 one target native speaker. A report should be done by the proofreader to explicit the improvements provided during the process.
- In addition to all the general translation criteria (mostly shared for the translation of different data types and domains), a number of criteria are defined to handle specific points in different tasks. For instance, some language directions have their own guidelines if there is a need to focus on specificities for any given language (e.g. proper names in Arabic) or at a larger and more complex scale, some data types, such as audio transcriptions, which have a full set of criteria to help translators in their work. The following serves to illustrate some of the issues aimed at in the guidelines for the translation of audio data:
- Reiterated words must not be translated, e.g. "la la Russie" should be translated into: "das Russland".
- Onomatopoeia such as "euh", "hmm", "ah", etc. in the source document (always preceded by a "&") must not appear in the translation.
- An isolated ampersand (&) in the source document indicates an unintelligible part of speech and must not be translated.
- Annotations contained between square brackets, e.g. "[b]", "[b-]", must not be translated.
- Some words are preceded by the * or "^^" symbols (for words which are mispronounced in the audio recording and words whose spelling is uncertain, respectively). Such words must be translated. However, the symbol before must not appear in the translation. For instance, "j*essayerais" should be translated into "Ich werde versuchen".

10. Quality Control: Validation

Among the different translation steps, the validation protocol plays an important role as it defines a series of points that need to be checked in order to guarantee good-quality output while bearing in mind the needs of each corpus.

10.1 Setting up the Validation Team

In order to assure the quality of the translations, ELDA has also developed specific validation guidelines to enforce the following policies for each translation delivery:

1. Recruitment of fluent bilinguals to control the translation quality. They validate the translations against the validation guidelines. Every delivery is subject to revision.

2. A subset of documents is randomly selected for quality control. The selected sample translations are then graded.

3. To ensure consistency from one review to another, an error typology has been adopted to grade translations. Each error is associated with a penalty point, which allows us to compute a validation score.

4. If reaching a defined threshold (level of errors), the translation is rejected and the whole delivery is sent back to the translation team for improvement.

5. If a delivery is sent back to the translation team for further proofreading, the improved version must be completed within an agreed time. This time will be established with regard to the number of words to be proofread.

10.2 Validation Criteria

The goal of the validation guidelines is to provide a methodology for validating the translations produced. These are adapted to the specificities of each data production project. For instance, although the translation error typology is similar to that of the TC- STAR project, both the penalty points applied and the validation scores established have been adapted to the needs of the Quaero project. Given the high-quality expectations of this project, the validation score threshold defined is very strict, only accepting 1 penalty point per 100 words (see Section "Validation by Human Experts" for further details on the penalty typology).

The resulting translations are thus divided into accepted and rejected. An accepted translation is kept. However, a rejected translation is sent back to the translation agency with a validation report. A delay is agreed upon for the return of a new translation. As the validation procedure is carried out on a sample of each translation, the new translation to be provided by the translation agency must not be a corrected version of this sample only, but of the full file.

The quality control of the data here described consists of both an automatic and a manual procedure. These are further detailed in the subsections below.

10.3 Automatic Validation

An automatic validation is provided when a translation is received by ELDA from the translation agency. If numerous and irrefutable errors are found, the translation is immediately sent back to the translation agency. The following issues are considered in this automatic validation:

1. A spell checker checks the translation automatically. If necessary, the spell checker is adapted to the corpus lexicon. The errors found are considered as lexical errors and are reported.

2. The format of the corpus is automatically validated too, checking whether the specifications established in the translation guidelines have been followed. The translation might be sent back to the translation agency if many errors are found.

3. In the case of the corpus with paraphrases, these variations are checked so as to ensure that translation repetitions have been avoided. According to the number of errors, specific sentences might be sent back to the translation agency, or the whole corpus.

Once the translation has passed this automatic validation stage, the data goes into the following human validation.

10.4 Validation by Human Experts

Regarding manual validation, as mentioned above, this takes place over a randomly selected sample of data. For each delivery, a random subset of the corpus is selected, until the number of words adds up to about 5-7% of the source text (considering full sentences) translated by a single translator. Then, the validation corpus is offered to the validators containing both source and target texts.

The validation task consists in proofreading the texts and whenever a problematic point arises:

- Labelling the problematic sentence (with a label from the list of problems detailed in Table 1);
- Proposing a correction/improvement, if possible, and/or a short explanation of the error found.

The aim of the validator is to evaluate if the translation is of good quality, not redo it, as when aiming at producing a final version of a document for publication. Such revision/correction is the task of the translation agency. However, since we are evaluating the quality of the data we often need validators to provide arguments (some corrections, comments) to justify the validator's criteria/decisions.

In order to perform some document validation, the following technical issues are taken into account:

1. The files to be validated are provided to validators in text format (or Microsoft Office Word, if required), a simple format that allows easy handling. Validators are expected to submit their files respecting this original format.

2. The sentences to be validated look as follows:

- source sentence
- translated sentence
- blank line

3. Corrections and notification of errors are provided per sentence. If no remark or correction is to be provided by the validator, this format remains the same. However, if a segment contains an error, then a new line is inserted starting with "#" right after the segment. After the "#" follows the type of error (5 categories, according to the scheme described below), together with the correction or indication of the error itself. The resulting format would be as follows :

- source sentence translated sentence
- # error type + correction or indication of the error blank line

In the case of multiple errors, each error is on a new line starting with "#".

4. To ensure consistency from one validator to another, the following system has been adopted for grading translations, following the translation error typology illustrated in Table 1. Validators use the following types/labels (whenever possible) to tag translation errors:

- *Syntactic* errors are those found in grammatical categories. These comprise errors such as problems with verb tense, co-reference and inflection. Furthermore, syntactic errors are also those where there has been a misinterpretation of the grammatical relationships among 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 their 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.

Table 1 : Translation error typology

Error type	Penalty score
Syntactic	3 points
Lexical	3 points
Poor usage of the target language	1 point
Capitalisation	1 point
Punctuation	½ point

5. In order to avoid over-correction or over-validation, it is essential that the given translation receives the benefit of the doubt in terms of quality. This means that only clear errors should be indicated. However, if validators have proposals for improvement, they are invited to provide them to us by labelling them as "Preference". The thin line between a Preference and an error is easily crossed as translators find it sometimes hard not to try and impose their preferred way of translating over a case that seems dubious to them.

6. When several translations (multiple references for MT evaluation) are produced for a same source text, these are validated separately, each of them going through the same validation procedure described above. However, serious errors (syntactic and lexical) detected on either of the translated texts are also verified in the other translations in order to avoid the proliferation of problematic cases. This verification among the different translations is carried out by ELDA, based on the results/findings of the validations.

Points are deducted according to both the number and type of errors found. If the number of points goes over a certain allowed threshold, the translation is rejected and, thus sent back to the translation agency for correction. Every time a new translation is validated, a validation report is created, allowing the follow-up of the translation procedure and the interaction between ELDA and the translation agency.

As mentioned earlier, the validation threshold is set up at 1 penalty point per 100 words. This means that if a syntactic or a lexical error is found (each of these error types is worth 3 points), this will be the only mistake allowed per 300 words of translated data. This is very strict as procedure, but it guarantees the achievement of very polished and high- quality data, as required in our evaluation context.

11. Validation is not a Trivial Task

A sometimes problematic issue is that quality control involves translators evaluating other translators without being a translation task as such. This is somehow tricky as some translators

struggle to reach a balance on what is needed during validation. We encounter the following situations:

- On the one hand, some translators do not want to do this kind of work as they are either not appealed by the task or they find it hard to do what they consider as scrutinising or criticising other colleagues' work. In both cases, they were trained to translate and/or proofread and validation stands out of their field of expertise. Furthermore, in the case of the latter, listing error types seems to bring up some sort of ethical issue that troubles them. This is the case of a few, but it happens.
- On the other hand, and among those translators who take over the role of validators, there are some who seem to feel compelled to over-correct the translated data, probably aiming to justify their role. This is one of those situations where following the work carried out closely, filtering the outcome and discussing with the teams is absolutely essential. Thanks to those numerous exchanges, we have managed to make a very good use of the "Preference" category we had mentioned in the previous section. This is also one of the reasons why working with experienced and trained professionals simplifies the task, as these are further knowledgeable also with regard to our needs and our expectations.

Moreover, validating the quality of translation may be far from straightforward, even for a translation expert. A large number of validations require an analysis of the issues raised. Indeed, validations may give raise to comments from the translation team, who have received the validation report (containing, among other things, both the error types detected in the validated translation sample and the score obtained) and have pointed out their disagreement over some particular points. In their opinion, some translations have been wrongly classed as mistakes while it was a "simple matter of translation "preference". In very extreme cases where none of the parties agrees to the other's opinion, a third expert may need to be called in to give his/her opinion. This needs to be cleared out as a validation report stating a failed validation enforces the translation team to correct the whole corpus taking into account the types of errors detected.

For instance, in the framework of one of our German-to-French productions of corpora derived from transcriptions (22k source words), the validation procedure has required two validations and thus one correction to reach the required quality level[10]. This does not take into account the efforts (time + manpower) spent either handling the disagreements between the translation team and the validation team, or carrying out the automatic format validations of the data resubmitted by the translation team.

12. How Long for a Full Translation and Validation Procedure until Delivery?

For the above-mentioned project and given the data size handled, the production effort could be quantitatively summarised in terms of duration. Once the translation team (translator and proofreader) and the validator(s) have been recruited[11], one such corpus requires about 50-60 working days of production time. These figures will be slightly increased if a full revision and correction of the corpus are required after the full validation. The timing can be broken down as follows:

1. First delivery: this comprises the first translated and proofread data (7 working days), which is sent for an initial quality control. An early detection of unexpected problems allows an easier management of corrections. Besides, such problems are often due to a misinterpretation of the guidelines. Thus, this early detection allows for an early clearing out and handling of unclear or ambiguous points.

(10) The number of validations performed per corpus produced is rarely higher than three for most of our produced corpora.
(11) When these are not already part of the regular working team, professionals are tested in order to join the project.

2. First validation: first delivery is validated (1-2 days, depending on delivery time).

3. Second delivery: comments from the first validation are to be taken into account to produce a second delivery half way through the project (10 days approximately).

4. Second validation: second delivery is validated (2-4 days, depending on data size and delivery time).

5. Final delivery: delivery of full data (25 days approximately).

6. Full validation (2-4 days, depending on data size and delivery time).

7. Data revision and correction (if necessary, according to validation results: 7-10 days approximately). If this is the case, a new validation will be required (step 8).

8. Final validation (2-4 days, depending on data size and delivery time). Should this validation fail, the data will be back to step 7 until the required quality is reached.

Further extra tasks and costs could also be incurred during data production. For instance, disagreements or questions during translation and validation represent an extra cost for the project in terms of translators/proofreaders and validators' time. For the translation team, this is part of their estimated cost as overhead (as they are meant to deliver high quality work), but in what regards validators, this represents an extra cost which is invoiced at the same price as their validation work (payment per word when sentences or texts are to be reconsidered).

12.1 Complexity in Producing Data for SLT

As discussed throughout the paper, a number of issues arise during the production of data for the evaluation of Machine Translation. The level of complexity is further increased as this production involves working on spontaneous speech transcriptions. The sections below look at this matter in detail.

12.2 Complexity Sources

Spontaneous speech language structure goes well beyond the scholarly learnt syntax. The day-to-day issues faced by the translators go certainly much further than the standard translation complexity.

A recurring issue encountered by the translators is often linked to the search for a balance between translation precision and fluidity. This is particularly problematic given the task, which consists in translating transcriptions which are spontaneous and grammatically fragmented by nature. The exercise of conveying meaning while respecting fluidity, is often controversial in the translation of sentences (segments) such as the following:

« et, là, c'est compliqué, il y a les lobbies.
% % qui est qui sont en qui se mettent en. »

Even the handling of speech phenomena, such as the following, is disturbing in the translation task:

« &hein , on on d() on peut prendre d ' autant plus de risques que la maladie qu ' on traite est plus grave , &hein . »

Moreover, translation choices do not represent universal truths with one single possible solution. This is a statement that can be applied to translation in general and very particularly to this type of data. This is one of the reasons behind the discussions and disagreements incurred by our validators and translators, mostly in terms of « wrong usage of target language » and « preference ». As we said earlier, the borderline between these categories is not always easy to establish, and indeed even less easy when approaching the translation of spontaneous speech.

In addition to this, limiting the choice of possibilities to the strictly necessary constitutes a real challenge for the translators producing data for the evaluation of technologies. The instructions that we provide for the translators, proofreaders and validators are not necessarily part of their professional background and formation. Their everyday work often contains a creativity factor that is generally refrained when producing evaluation corpora. Furthermore, even the mere fact of having to translate from a source with mistakes, disfluencies or incoherences is often confusing.

As a result of all this, it needs to be said that it is with the help of the close collaboration carried out by translators and proofreaders, who work together towards a joint output, that the difficulties faced during the project are overcome. The translation teams report on troublesome cases and decisions taken whenever these are questioned, and they keep up a follow-up of their internal discussions. This is very helpful for both a) them, in order to deliver a product, and b) us, to be able to discern what has been done when (negative or doubtful) feedback from validators is received.

Last but not least, a reference should be made to the segmentation of the speech data, problem which has been already described in the section « Resegmentation of the Source Data », but of relevance at this point. Let us remember that the time-based segmentation, traditionally used in Automatic Speech Recognition (ASR) and independent of the semantic units is a major issue for translators. Such segmentation implies, for instance, that units (sentences) may be split when breathing, or just strategically stopped when the speaker (typical in broadcasting) tries to avoid being interrupted and to keep people's attention. This may not look so complicated at first sight, but we should remember that the data produced needs to be aligned with their sources for evaluation purposes. Such alignment will become even more challenging when the syntax between the source and the target languages is very different. This is so for a number of production projects that have taken place and where language pairs such as French and German are under consideration. A French source sentence may be divided into two or more segments that need to be translated into German, which will pose a problem for translators who are obliged to preserve the segmentation of units that may be impossible to match/align afterwards. In order to handle this difficulty, the re-segmentation of the source data may be foreseen. A re-segmentation expert takes care of separating or merging segments in order to obtain well-formed and self-contained sentences.

12.3 Particularly Difficult Problems

A number of phenomena which are specific to speech data are the cause for major complication during translation. Here follow some of them:

- The difficulty to understand transcribed data has provoked a lot of discussions since translators have had to face either non-understandable source text or incomplete sentences. For instance, in one occasion the translator did not understand what a speaker wanted to say. This was the case for several of this speaker's utterances, which obliged the translator to interpret the transcription to proceed with translation.
- Transcription errors (like spelling errors, missing words, etc.) disturb the translators, proofreaders and validators:

- A translator detected a potential transcription problem with "das sieht" ("cela voit", this sees) at the end of a segment. He estimated, according to the rest of the source sentence, that it should rather be "das sind" ("ce sont", these are), which fits very well with the following segment "dreizehn Prozent des Bruttoinlandsprodukts" ("[cela représente] treize pour cent du produit int´erieur brut", [this represents] thirteen percent of the gross domestic product).
- The transcription "bloß in Zentralamerika sowie in Zentralamerika sind sie ganz stark vertreten." makes no sense on its own and, by listening to the audio document, one should actually hear "plus in Zentralamerika. In Zentralamerika sind sie ganz stark vertreten"..

In all these cases, making the audio available for the translators to use as reference is crucial in order to allow them understand the transcription. This helps them to either find or check the words to be translated and also to disambiguate problematic cases.

▶ The difficulty in understanding or interpreting the translation guidelines, in particular when translators need to deal with two different actions at the same time. For instance, this is the case of the following points:

- repeated words that must be translated only once (for instance, "la la Russie" is to be translated into "das Russland") and words that are partially pronounced and should be transcribed using the "-" symbol and tagged with the "%pw" tag in their translation (for instance, "wir werden n- eine pfanne nehmen" is translated as "nous allons prendre%pw une poêle", i.e. we will take a pan).

Further to the speech-nature related obstacles just mentioned, two more types can be added:

▶ The difficulty in establishing a balance between a translation that is close to the source text (adequacy) and a fluent output in the target language. This is often a source of disagreement among all those involved in the production chain (translators, proofreaders and validators). Translators and proofreaders try to remain as faithful to the source and to the guidelines as possible, sometimes sacrifying fluidity slightly. In this case, they may be reprimanded during validation. This may also do the opposite, with validators penalising a certain unwanted creativity. In either case, translations are looked at on a one-to-one basis and decisions are taken accordingly, always with the closest respect for the guidelines. Below follow a couple of such problematic cases:

- The speaker talks about a particularly expensive thing by using the term "Herzkreislaufbehandlung" and the translator has interpreted it as referring to a "transplantation cardiaque" (i.e. heart transplant). However, since he is supposed to translate only what it is said without any interpretation, he has chosen a more literal translation with a "traitement cardiovasculaire simple" (i.e. simple cardiovascular treatment).
- In the translation of "Sie sind einer der Mitverfasser des Drogenberichtes", the translator used "Vous êtes l'un des coauteurs..." (i.e. You are one of the coauthors...) instead of what would have been his preferred choice (more creative and not 100% literal) "Vous avez participé à la rédaction du rapport sur la drogue" (i.e You participated in the writing-up of the drug report). This was done with the aim of keeping the original structure of the sentence, and thus following the translation guidelines.

▶ Knowledge about context is essential for certain translations, which is achieved with the help of the audio data that go with the transcriptions. However, it should mentioned that for some segments, it is the visual information that would help disambiguate the transcription. Since this was not available, the translators remained as literal and close to the source sentence as possible, as shown in the following example:

- The segment "der Bonner Wahlkreisabgeordnete Westerwelle drückt sogar den Knopf." ("Le député de la circonscription électorale de Bonn Westerwelle appuie même sur le bouton", i.e. The representative from the electoral district of Bonn, Westerwelle, even presses the button) is ambiguous and it seems that, in the audio document, somebody is taking pictures. Thus, if the speaker talks about a camera, "bouton"

(button) should be replaced by "déclencheur" (kind of trigger) in French. On the other hand, he could very well be talking about an "interrupteur" (switch). This is impossible to tell without actual access to the video data.

13. Lessons Learned and Adopted Solutions

The lessons learnt throughout these years are numerous, both in technical terms and in management. However, the first message to pass on to all those interested in such data production initiatives is that of great respect for the work of the language professionals behind such production. We are well aware that such a procedure has a cost, however, the high quality obtained by human translators is not negligible, and neither is the fact that even if somehow costly, such data are reusable and shareable, which is part of the sustainability plan for such effort. This allows for a future use of quality reference data in further campaigns or evaluation activities.

Regarding the technicalities that can be improved within the project, there is always a few of them as a further project concludes. It is our aim to provide rich and updated guidelines, yet, no matter how many cases are covered by the existing documents, the exception will always pop up. Points as those discussed in the previous section are always a guide to improve our guidelines. Whenever an ambiguity or a non-covered case is pinpointed, this is studied and included in the next version of the guidelines. This has helped us define solutions for specific problems, such as the translation of titles for movies, TV series, broadcast programmes and books. No specific point had been defined to handle this in our translation guidelines and we have had to manage it a posteriori. The problem was raised when detecting that two different translators were producing different translations for the same TV series title ("Mike Nelson Abenteuer unter Wasser"): one of the translators had translated it literaly ("Les aventures de Mike Nelson sous l'eau", i.e. The adventures of Mike Nelson under water), while the other one provided a standardised name ("Remous", i.e. See hunt). As a consequence, it was decided that translators should try first to find already standardized translations, and otherwise, they could leave the titles in their source language if no standardized version was found. This is in fact the usual procedure in the professional translation world.

Certain errors found in the source texts have been the cause of some deep consideration. Those cases have been discussed with the translation teams so as to come up with appropriate solutions according to the specifications and without incurring into project delays. As we have seen earlier in the document, some of the errors found concerned transcription mistakes that whenever causing translators to be stuck, were resolved with the help of the audio data. Even if stated otherwise in the guidelines, the errors reported may concern minor format and encoding issues in the target reference data produced. Given that translators make use of a variety of text editors and translation tools to help them in their work, a post-processing step keeps an open eye for anything unwanted.

In terms of translation quality, a considerable number of errors during validation penalize what we call "wrong usage of the target language". There have been many corrections and discussions concerning this matter, which as the reader may imagine, is a good source of disagreement. Some of these discussions argue between the "error" and the "preference" concepts. An agile communication with the teams and trying to clear and define things from the very beginning, or as early as the issues are raised, is essential for the success of the project, where success refers to output quality while not exceeding costs and timing.

In this regard, early deliveries and validations are a must: regardless of collaborating with well-known professionals, each new project has its specificities. As seen in the quality control section, an early validation after a few days of work guarantees an early control of unforeseen problems. Such early detection allows for a correction of little data and a re- definition of the

detected problems. It is a much lighter procedure than aiming to change and correct once the production is well advanced.

Despite all the possible controls to be carried out, it should be said that a perfect 100% error-free translation (for such data size and features) does not exist. Translators are human and there are a number of parameters (such as tiredness, lack of concentration, etc.) which play an important role. Therefore, quality cannot improve further after a number of revisions. If the quality achieved is not up to our satisfaction, we should reconsider whether our expectations are realistic. If they are and we are still not satisfied, it may be that the team is not really adapted to the task. In that case, we should not hesitate to change the team and the quicker we take action, the less the project will be suffering from it.

Last but not least, one of the simplest and most valuable pieces of advice is to try and keep experienced translators/proofreaders and validators as part of our team. Experience tells us that all the unexpected may take place, so if we can lower the probabilities, the production will benefit from it. Working with language experts who perform and interact well is a big advantage for the project and its members.

Conclusions

Parallel corpus production by means of manual translations is a complex and demanding task, in particular when aiming at the production of reference evaluation data. This paper aims at providing an overview of the procedures and guidelines developed for that purpose as well as of the large complexity behind the task. For that purpose, we capitalize on ELDA's expertise and experience throughout years of data production and, very particularly, for the production projects that have taken place within Quaero.

The procedures detailed comprise data preparation, translation and validation teams setting up, translation and quality control, as well as translation and validation guidelines definition. We aim at providing a matured account of such procedures which should guide data producers to either endeavor in data production or in any of their required steps.

Moreover, we also describe in detail the complexity of the data production experience concerning spontaneous speech. Analysing the problems encountered and looking at the solutions adopted has helped us improve both the outcome of the project(s) and the full procedure as a whole. Documents and protocols have been consequently updated to meet the latest needs of the task, following the lessons learnt.

Acknowledgements

This work has been financed by the Institute for Multilingual and Multimedia Information (IMMI) in the framework of the Quaero project. We would also like to thank all the translators, proofreaders and validators, for their big efforts to meet our production needs and for their eagerness to explore new adventures in translation.

Bibliography

Banerjee, S. and Lavie, A. (2005), "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments". In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, June 2005, Ann Arbor, Michigan, U.S., pp. 65-72.

Callison-Burch, C., Koehn, P., Monz, C., Peterson, K., Przybocki M. and Zaidan, O.F. (2010), "Findings of the 2010 Joint Workshop on Statistical Machine Translation and Metrics for Machine Translation". In Proceedings of the Fifth Workshop on Statistical

Machine Translation, July 2010, Uppsala, Sweden, pp. 10-17.

Doddington, G. (2002), "Automatic Evaluation of Machine Translation Quality Using N- gram Co-occurrence Statistics". In Proceedings of the Second International Conference on Human Language Technology, pp.138-145.

Hamon, O., Choukri, K. (2011), "Evaluation Methodology and Results for English-to- Arabic MT". In Proceedings of MT Summit XIII, September 2011, Xiamen, China, pp. 480-487.

Hamon, O., Hartley, A., Popescu-Belis, A., Choukri, K. (2007). "Assessing Human and Automated Quality Judgments in the French MT Evaluation Campaign CESTA". In Proceedings of MT Summit XI, September 2007, Copenhagen, Denmark.

Koehn, P. (2005), "Europarl: A Parallel Corpus for Statistical Machine Translation". In Proceedings of the Tenth Machine Translation Summit, 2005, Phuket, Thailand, pp. 79–86.

Mostefa, D., Hamon, O., Moreau, N., Choukri, K. (2007), "Technological Showcase and End-to-End Evaluation Architecture", Technology and Corpora for Speech to Speech Translation (TC-STAR) project. Deliverable D30, May 2007.

Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J. (2001), "BLEU: a Method for Automatic Evaluation of Machine Translation". IBM Research Division, Thomas J. Watson Research Center, Technical Report 2001.