Building Reference Data for MT Evaluation

Victoria Arranz

ELDA
Paris, France
arranz@elda.org
http://www.elda.org/ or http://www.elra.info/

Overview

- Introductory Issues
- Protocol for the Production of Parallel Corpora for MT Evaluation
- Translation and Proofreading Protocol
- Quality Control Protocol
- But what Happens when Speech Complexity Comes along?
- Concluding Remarks

What is Reference Data?

- Data used to « automatically » measure system output

- Also known as *test* data

Questions behind Reference Data

BUT:

- How reliable are these data?
- Are they comparable to human evaluation?
- How much data do we need? (Size, number of references)
- What quality/error level do we need / can we afford in the data?

How do we Build such Data?

- Either manually done or with a large manual component
- Test data: higher quality than training or development data
- If initial automatic component (eg. Crawling?): manual revision and correction is a must → PANACEA

How do we Build such Data?

- Moreover, unexpected issues: even crawled data needs some pre-crawling manual work:
  - Technically: customizing crawlers, not that trivial if targetting specific domains (focused-crawlers?) and clean data (boilerplate removal,…)
  - Logistically: we just cannot get anyone's data from internet ☹ → Intellectual Property Rights (IPR)
  - IPR issues: need to be checked and if need be, negotiations done

How do we Build such Data?

- Once rights have been cleared for the data to be used: what are the steps to build MT/SLT evaluation reference data?

To be continued… but before….

Production Context

- ELDA and its production these past few years for the Quaero evaluation campaigns
- Initially based on TC-STAR and GALE experience
- Improvements over the years:
  - To handle non treated or unclear points
  - To take into account speech-related phenomena: disfluencies, such as onomatopeia or partially-pronounced or reiterated words.

What Kind of Data do we Build?

- Type and domain:
  - Text (e.g., journalistic; pharmacology patents)
  - Audio data (e.g., radio and TV transcriptions, debates, parliamentary speeches)
- Languages:
  - Lately → Quaero campaigns: Arabic, Chinese, English, French, German.
  - Also: more exotic languages like Pashto → complex to find language experts

What Kind of Data do we Build?

- Data sizes:
  - Between 10/15K to the 22/27K words in Quaero
  - Smaller than training data
  - But higher quality
  - Size is restricted by usage as well as cost
- Cost:
  - Quality has a cost
  - Management has a cost

How long to Produce the Data?

For about 22K word source:

- Average of 50/60 working days for full translation procedure (with full cycle till a full validation).
- If validation fails and data requires revision, timing will vary.
- If speech sources need to be corrected or re-segmented: further 10/15 working days.

Protocol for the Production of Parallel Corpora for MT Evaluation

Protocol Steps

1. **Translation** to be done by a bilingual translator whose mother tongue is the target language
2. **Proofreading**, corrections and homogenization to be done by a native speaker of the target language
3. **Automatic validation** of both format and content
4. **Manual validation** by an expert in translation and proofreading
5. Production of a **validation report**
6. *If the corpus is rejected, go back to step 1 on the basis of the validation report*

Translation and Proofreading Protocol

Translation Team Setting-Up

- Clearly establish the number of translations (references) so that teams do not overlap

- Each translation should be done by a different translation team

Translation Team Setting-Up

- Each team might be composed of:
  - A bilingual translator, native speaker of the target language, who will be in charge of one of the translations required per corpus
  - A target native speaker bilingual who proofreads and edits the output of the translator. (S)he is also in charge of the homogenisation of the whole corpus, especially regarding the vocabulary

Translation Team Setting-Up

- Notice that the translations must be systematically finalised and checked by a target native speaker
- The translation team should not change during the course of translation, and the team must be fully documented:
  - Name and expertise details of team members
  - Team composition details and order of file processing

** Teams become « established experts » over the years for specific languages and domains*

Source Data Description

- Definition of domain, format, encoding, DTD and all necessary pieces of information. Eg. :
  - **Data type: broadcast news and parliamentary speeches**
  - **Encoding (UTF-8 and pseudo XML), docid attributes**
  - **Source data: segmented using the time-based segmentation**
  - **Each time-based segment is identified with a ”<seg id=”...”>**
  - **Encoding to render the target files**

Data Production Sample

Source: Chinese patents in pharmacology

- <?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE mteval SYSTEM "ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-xml-v1.3.dtd">
<mteval>
<srcset setid="raw data collection_zh-en" srclang="Chinese">
<doc docid="CN200910003515" genre="patent">
<seg id="6" >一种蒙药材草乌的炮制方法</seg>
<seg id="7" >本发明公开了一种蒙药材草乌的炮制方法，该方法在常温下将蒙药草乌生品浸在水中润至内无干心，切成５～１０ｍｍ的厚片，在９０～１０５℃的条件下烘制４～１０小时，即得草乌炮制品。</seg>
<seg id="8" >该方法使草乌炮制品质量易于控制，药效组分损失较小，起到了减毒增效的作用，同时炮制方法简便易行，适合工业化大规模生产。</seg>
</doc>
</srcset>
</mteval>

Data Production Sample

Target: English patents in pharmacology

- <?xml version="1.0" encoding="UTF-8"?>
  <!DOCTYPE mteval SYSTEM "ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-xml-v1.3.dtd">
  <mteval>
  <srcset setid="raw data collection_zh-en" srclang="Chinese">
  <doc docid="CN200910003515" genre="patent">
  <seg id="6" > A Method for Refining the Mongolian Medicinal Substance Radix Aconiti Agrestis</seg>
  <seg id="7" > The invention presents a method for refining Radix Aconiti Agrestis, a Mongolian medicinal substance. The method consists of soaking the raw Mongolia Radix Aconiti Agrestis in room-temperature water until fully saturated, slicing it into pieces 5-10 mm thick, then baking it for 4-10 hours at 90-105 ºC to obtain the final drug substance.</seg>
  <seg id="8" > This method results in better quality control of the refined Radix Aconiti Agrestis. Because less of the active components are lost during the process, the yielded substance is a more effective detoxicant. At the same time, the refining method is simple and easy to implement, making it suitable for large-scale industrial production.</seg>
  </doc>
  </srcset>
  </mteval>

Translation Guidelines

- Most crucial part of the information to share with the translation team
- Even if best practices are used by translators, specific points need to be cleared out in advance
- This is very important when particularly complex data are handled, such as speech transcriptions

Translation Guidelines

- Even with all these clarifications:
  - Ambigüity takes place
  - Misunderstanding is a threat: how fluent to be fluency acceptable??? ☹
  - Multiple discussions may take place during the project…
  - …and these issues are particularly detected during the 1st validations
  - Thus: early validations are a must!!!!

Translation Guidelines

Some specific guidelines…

…some very hard to accomplish:

- The target translation must be **faithful** to the original source text **in terms of meaning and style**. […] this should be achieved without sacrifying grammaticality, fluency and naturalness. That is to say, **being faithful does not mean producing literal translations**.

Translation Guidelines

- The **tone and register** of the language should be respected. […] this state of mind should be also expressed in the target language, conveying the same tone.

- The translation should be as **factual** as possible, […] **without adding/removing information**.

Translation Guidelines

Particularly important for technology evaluation:

- 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.

Translation Guidelines

Some are part of the guidelines evolution:

- 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.

Translation Guidelines

Guidelines are adapted to data specificities, such as language:

- Regarding proper names, […] **in the case of Arabic, this may imply providing a different translation from that suggested in Modern Arabic.**

Translation Guidelines

Some are specific to cover data type and handle, for instance, transcription data:

- The style (i.e. oral transcription) of the source document must be kept in the translation
- Reiterated words must not be translated, […]
- Onomatopoeia such as ”euh”, ”hmm”, ”ah”, […] must not appear in the translation
- Unintelligible parts of speech […]

Translation Guidelines

- Annotations contained between square brackets […]
- Mispronounced words whose spelling is uncertain […]

- Partial words must be annotated with the ”%pw” tag, […]

  E.g.:

  the segment ”wir werden jetzt n- eine pfanne nehmen...” must be translated into ”Nous allons maintenant prendre %pw une poêle...”.

Quality Control Protocol

Quality Control

- Crucial in the data production process
- Quality control is carried out by:
  - Means of a pre-defined procedure: validation guidelines
  - Validation experts: tested and formed for that particular task
- Validation task → not simple:
  - It is not a translation task
  - It involves evaluating other language and translation professionals

Quality Control

- ELDA hires fluent bilinguals to control the translation quality. They validate the translations against the translation guidelines provided to the translation team
- Every delivery is subject to this revision
- For each delivery, we randomly select a subset of the documents. The selected sample translation is then graded
- To ensure consistency from one review to another, a system has been adopted for grading translations

Quality Control

Current translation error typology:

Error type	Penalty score
Syntactic	3 points
Lexical	3 points
Wrong usage of the target language	1 point
Uppercase or orthographic error	1 point
Punctuation	½ point (max. of 10 points)

Quality Control

- If reaching a defined level of errors, the translation is rejected and the whole delivery is sent back to the translation team for improvement

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

Validation Guidelines

- Procedure for quality control within a specific context and under specific conditions
- A randomly chosen sample of 5% of translated corpus is used
- To reach high quality translations: very strict validation
- Very few errors are accepted → project and data purpose dependant

- However: the « perfect, 100% error-free » translation does not exist
- After a number of validations: quality cannot improve any further → translators and proofreaders do not see the errors any longer

Validation Guidelines

- Quaero: 1 penalty point per 100 words
- This implies that only 1 lexical error is accepted per 300 words
- Clear cutting between errors and preferences is not straightforward
- Validation points may be discussed with validators and translation teams
- Sometimes project timing suffers from disagreements between validators' decisions and translators

Validation Guidelines

- These guidelines provide:
  - Detailed information + guidelines on translation: validators have access to all information translators are provided with
  - Details on automatic validation carried out by ELDA
  - Details on manual validation to be carried out by human experts

Automatic Validation

- Spell checking: If necessary, adapted to the corpus lexicon
- The format of the corpus is automatically validated too, checking whether the specifications have been followed
- In the case of the corpus with paraphrases, these variations are checked so as to ensure that translation repetitions have been avoided

Validation by Human Experts

- The validation task consists in proofreading the texts and whenever a problematic point arises:
  - Label the problematic sentence (with a label from the list of problems detailed in error typology)
  - Propose a correction/improvement for the problematic part, if possible and/or a short explanation of the error found

Validation by Human Experts

- Validators are given specifications on:
  - Format of files to be validated (generally txt)
  - Internal format of file content (which line with what)
  - How to indicate errors and comments
  - Full definition of what each error type means. E.g.: « *Poor usage of target language* means awkward, unidiomatic usage of the target language and failure to use commonly recognised titles and terms. »
  - Translations should receive the benefit of the doubt
  - Different translations of a same source are validated separately, but serious errors found in one are checked in the others

Validation Report

- A validation report is produced for every validation
- It allows the follow-up of the translation procedure and interaction between ELDA and the translation team
- It provides:
  - Description of translation sample (# of words)
  - Details on errors found
  - Conclusions reached: data accepted or rejected

But what Happens when Speech Complexity Comes along?

Complexity in SLT

It may « *JUST* » imply pre-processing the source data before translation, for instance…

… re-segmenting + re-formatting for translation…

Re-segmentation Protocol

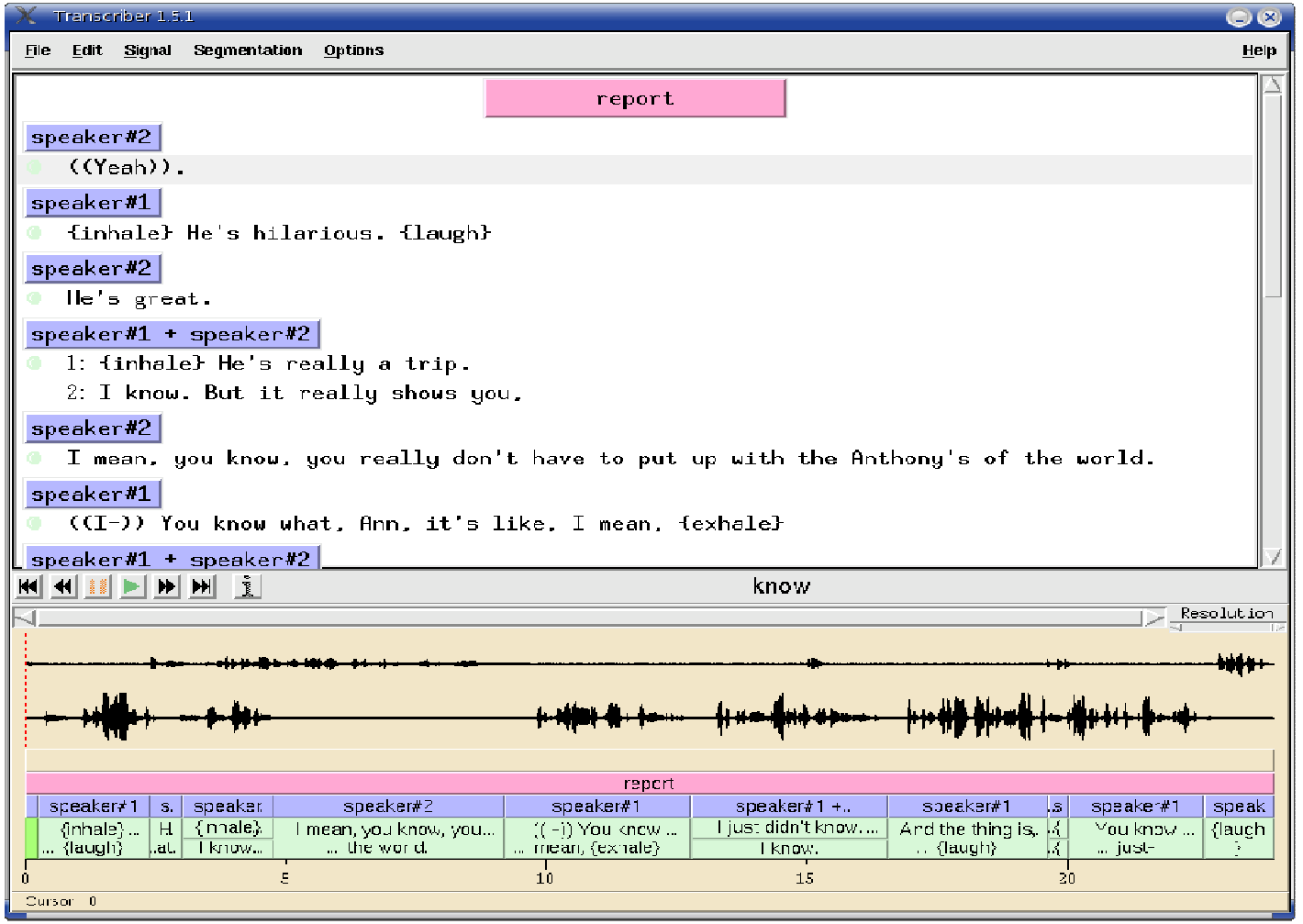
Re-segmentation

- Objective:
  - To prepare transcription data for translation
  - To obtain well-formed and self-contained sentences
- If re-segmentation needed:
  - A segmentation team needs to be put into place, comprising: segmenter(s) + validators

Re-segmentation

- Segmenter(s)' + validators' skills:
  - Native or native-like proficiency of the required language
  - Knowledge of linguistics
  - Well acquainted with the tools: Transcriber

Transcriber



Source: http://trans.sourceforge.net/en/screenshots.php

Re-segmentation

- Task consists in 2 actions:
  - Separating segments: when we find that several sentences have been wrongly placed within the same segment.
  - Merging segments: when we find that either a) one sentence has been split over more than one line or segment, or b) one sentence is too small to remain on its own.

- Specific guidelines are produced to guide the segmenters and validators throughout all specific speech phenomena they may encounter

Data Sample: Source

</Turn>

</Section>

<Section type="report" startTime="6.923" endTime="738.586">

<Turn speaker="spk1" startTime="6.923" endTime="43.275">

<Sync time="6.923"/>

à l' approche du scrutin , les candidats examinent de près les sondages . l' Alsace restera-t-elle à droite ? quid de la Corse ? de la Réunion ? de la Guyane ? nous ferons le point dans un instant .

<Event desc="b" type="noise" extent="instantaneous"/>

Data Sample: Target

<seg id="1" speaker="spk1" start="6.923" end="16.321">à l' approche du scrutin , les candidats examinent de près les sondages . l' Alsace restera-t-elle à droite ? quid de la Corse ? de la Réunion ? de la Guyane ? nous ferons le point dans un instant .</seg>

Data Sample: Target

All file ID information is still preserved, but content to be translated is simplified for translators and data users:

```
<?xml version="1.0" encoding="UTF-8"?>
  <!DOCTYPE mteval SYSTEM
  "ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-xml-
  v1.3.dtd">
  <mteval>
  <srcset setid="transcriptions" srclang="French">
  <doc
  docid="QRBC_FRE_FR_20100319_070000_FCULT_NE
  WS7H_POD.trs" genre="speech">
```

Complexity in SLT

Or it may imply dealing in translation with the added complexity of spontaneous speech…

☹ ☹ ☹

Complexity in SLT

- Spontaneous speech is well known for showing a side of language structure which goes well beyond the scholarly learnt syntax

- The day-to-day issues encountered by the translators go certainly much further than the standard translation complexity…

Particularly Difficult Problems

Specific to Speech Data:

- *The time-based segmentation*, traditionally used in ASR, is **independent of the semantic units** (e.g. units can be split when breathing)
- Even worse when the syntax between the source and the target language are very different (e.g. French and German):

<seg id="1"> C'est important euh pour euh les jeunes ici présents **de pouvoir** euh</seg>

<seg id="2"> **rencontrer** des hommes et femmes politiques mais c'est important aussi euh pour nous</seg>

Particularly Difficult Problems

- *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 certain translations, listening to the source becomes essential to allow translators understand the transcription

Particularly Difficult Problems

- *Transcription errors (*like spelling errors, missing words…*)* disturb the translators, proofreaders and validators
- Making the audio data available for the translators to use them as reference is crucial
- This helps them to find the words to be translated and also to disambiguate problematic cases

Particularly Difficult Problems

- *Difficulty in understanding or interpreting the translation guidelines*, in particular when translators need to deal with two different guideline points at the same time:
  - 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).

Particularly Difficult Problems

Other Encountered Problems

- *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*
- *Knowledge about context* is essential for certain translations, which is achieved with the help of the audio data that go with the transcriptions.

Concluding Remarks

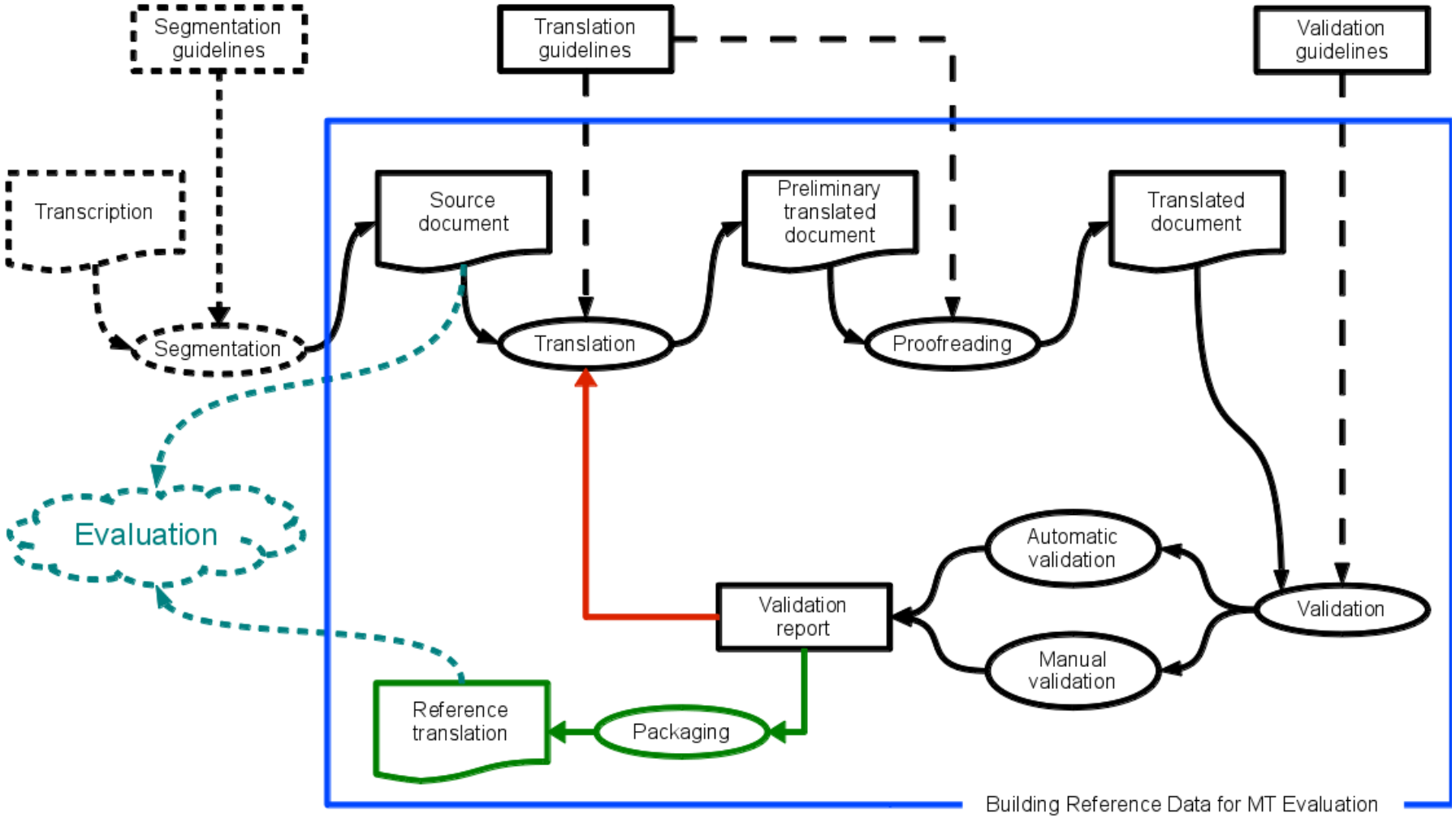
- Reference data:
  - Crucial to measure system output “automatically”
  - Important for evaluation reproduction and system comparison → together with Evaluation Packages
  - High quality: strict protocols
  - Manual / semi-manual production
  - Somehow costly, but reusable and shareable ☺
  - Task: a real challenge for both translation professionals and us, in particular if handling speech data

Concluding Remarks

- Protocols for:
  - Full production (with specifications till delivery to customer/ data user)
  - Translation + proofreading
  - Quality Control
  - Spontaneous speech data pre-processing

- Added complexity due to nature of speech data

Full Production Workflow



Acknowledgements

- External collaborators for the Quaero project:
  - Karim Boudahmane (DGA)
  - Martine Garnier-Rizet (IMMI)
  - And all those language specialists behind the scenes...
- Internal project collaborators:
  - Olivier Hamon
  - Hélène Mazo
  - Priscille Schneller
  - Kata Gabor
  - Jérémy Leixa

Victoria Arranz, Olivier Hamon, Karim Boudahmane, Martine Garnier-Rizet: *Protocol and Lessons Learnt from the Production of Parallel Corpora for the Evaluation of Spoken Language Translation.* IWSLT'2011, San Francisco, USA, Dec. 2011.

Thank you for your attention