Protocol and Lessons Learnt from the Production of Parallel Corpora for the Evaluation of Speech Translation Systems

Victoria Arranz[1], *Olivier Hamon*[1], *Karim Boudahmane*[2], *Martine Garnier-Rizet*[3]

[1]ELDA - 55-57, rue Brillat-Savarin, 75013 Paris, France
[2]DGA - 7-9 rue des Mathurins, 92221 Bagneux Cedex, France
[3]IMMI - Bât. 508, rue John von Neumann - Univ. Paris-Sud, BP133 - 91403 Orsay Cedex, France
{arranz,hamon}@elda.org, karim.boudahmane@dga.defense.gouv.fr, garnier@immi-labs.org

Abstract

Machine translation evaluation campaigns require the production of reference corpora to automatically measure system output. This paper describes recent efforts to create such data with the objective of measuring the quality of the systems participating in the Quaero evaluations. In particular, we focus on the protocols behind such production as well as all the issues raised by the complexity of the transcription data handled.

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. To a certain extent, best-known automatic evaluation measures, such as BLEU [1] or METEOR [2], together with most of the current measures [3], provide a comparison between reference translations and machine translations.

On the one hand, it is common to use a minimal set of two reference translations to estimate the quality of the evaluated machine output. A larger number of references offers a wider variety of language combinations, which is closer to the reality. In addition, this is potentially fairer for the systems under evaluation as it increases the possibility of matching the translated documents. Due to the numerous translation options, evaluation organizers have used up to 16 reference translations [4] in this attempt to carry out a fairer quality measure.

On the other hand, some evaluations are carried out using only one reference translation, with the risk of restricting the results to a single possible translation.

Bearing this in mind, the evaluator requires quality human translations that respect strict constraints regarding both the context of the source documents and the translation task. Corpus production may be done either with source texts translated by professionals, or by collecting documents that are already translated [5], for instance, from the Internet, and then performing a sentence-level alignment. Both methods have advantages and drawbacks. However, despite the higher cost for the former, it allows to obtain a more reliable translation of the source text given that the translator is well acquainted with the translation context and domain. Besides, a further issue of concern when collecting bilingual data, in particular from the Internet, is that of obtaining comparable rather than parallel corpora.

Furthermore, the translation of a source corpus is not limited to this one task when aiming to achieve high-quality data. Several steps are required for that purpose which follow a well established order: translation, proofreading, validation, correction or post-treatment (such as the alignment of the resulting data). Prior to this, translation and validation guidelines are produced and modified according to the translation direction and context. All these steps are a necessary evil in order to meet the expectations of MT actors as well as renew language resources (LRs) either for the evaluation or the training of MT systems.

In the last few years, several initiatives have emerged with the objective of producing reference translation corpora for evaluation. We have, for instance, the DARPA GALE programme [6] or the NIST Open MT series [7] for data produced by the LDC[1] (Linguistic Data Consortium) [8]. IWSLT evaluation campaigns [9] are also an example of parallel corpus production, in this case for the evaluation of speech translation systems.

For some years now, ELDA[2] (Evaluations and Language resources Distribution Agency) has also been responsible for the production of parallel corpora to test MT systems. Several translation directions and domains have been tackled during evaluation campaigns such as TC-STAR [10], CESTA [11] or MEDAR [12]. In collaboration with other more confidential production projects, the campaigns have allowed us to refine our translation and validation guidelines throughout the years, as well as our expertise in that domain.

This paper describes the production of parallel corpora carried out by ELDA for and in collaboration with the IMMI[3] (Institute for Multilingual and Multimedia Information) and

[1] http://www.ldc.upenn.edu
[2] http://www.elda.org
[3] http://www.immi-labs.org

the DGA[4] (Direction générale de l'armement), in the framework of the Quaero project. The paper opens with the description of the specifications followed for such corpus production, starting with the translation phase and then moving to the validation phase. Then, an anlysis of the issues encountered during the different steps of the production is presented. Such difficulties are mostly directly linked to the complexity of the source data. We focus here on the aspects concerning the speech data, which have represented the most important challenge. Finally, a list of improvements is provided, which have been implemented and have allowed us to achieve the level of quality required for the task.

2. Corpus Development

Corpus development is based on a predefined protocol, adapted to the specific needs of each corpus, that is, according to the nature of the data, its domain, the translation direction and the amount of data. Our protocol establishes the production conditions for the following phases: translation, proofreading, validation, correction and post-treatments that are performed during the data checking step.

In the context of our work, source data is provided by the partners of the Quaero project and already well formatted in XML. This allows an easy and correct alignment of source and target sentences. The paper focuses on the production of German-French parallel corpora for a 22K word corpus whose source is German audio transcriptions from broadcast news. Such transcription data is particularly complicated to handle for the translators, which increases the general complexity of the project and raises a number of issues. Two reference translations are produced that are then used for the evaluation of MT systems in the Quaero evaluation campaign.

Overall, six evaluation corpora of this kind (three German-to-French and three French-to-German) have been produced for the evaluation of speech MT within Quaero in 2009, 2010 and 2011. Each year, the evaluation sets from previous years have been used as development data. From the very beginning, sustainability has been one of the principles behind the production of these language resources, both in terms of quality and data availability.

For that reason, a strict protocol with rigorous revision has been set up for the creation of long-term LRs. As it has been pointed out at different occasions, the evaluation of current MT systems does not seem to improve significantly when evaluation data are of very high quality. However, this seems to be associated to the fact that such systems have been trained and developed with middle-quality data to start with. If higher-quality data are then made available for development, systems are expected to perform better and thus react better to the quality of the evaluation data.

3. Protocol for Corpus Production

The protocol to produce a translation corpus follows the following steps:

1. Translation to be done by a bilingual translator whose mother tongue is the target language.

2. Proofreading and corrections to be done by a proofreader whose mother tongue is the target language. (S)he will be in charge of homogenizing the result, when needed, in particular regarding the terminology used.

3. Automatic validation of both format and content.

4. Manual validation by an expert in translation and proofreading, who is bilingual and whose mother tongue is the target language.

5. Production of a validation report.

6. When the corpus is rejected, we go back to step 1 on the basis of the validation report.

These steps are based on those used within TC-STAR[5], but their implementation has been improved during the current project in order to take into account speech-related disfluencies, such as onomatopoeia or partially-pronounced or reiterated words.

In addition to the procedure described, translation guidelines are provided to the translation team, which is made up of a translator and a proofreader. These documents are meant to serve as guides in the translation task and to clarify specific points that may be prone to ambiguity or confusion. Translation guidelines are adapted following the features and specific needs of each corpus to be developed.

Like the translation team, validators are also given validation guidelines, which specify all the points to be considered as well as define how to point out and label any error they may find in the translation. This and other details on the quality control are further explained in the following section.

This protocol ensures that each corpus is produced by a team that should remain unchanged during the whole production work, which helps guarantee translation consistency. Despite all this, the goal of homogenization may remain crucial (especially for certain types of very domain specific data), since a single translator may, from time to time, also incur into inconsistency issues with the terms used. This makes proofreading by a different expert a key step not to miss. However, for the work described within this paper, we decided not to homogenize the two reference translations produced for a same source corpus. On the contrary, we thought that the language diversity produced by two different translations done by two different translators would contribute towards the language variety required for the evaluation of MT systems.

[4] http://www.defense.gouv.fr/dga/

[5] http://www.tcstar.org/

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	1/2 point (max. of 10 points)

Table 1: Translation error typology.

4. Quality Control

Among the different translation steps, the validation protocol plays an important role as it defines a series of points to be checked in order to guarantee a good-quality output while bearing in mind the needs of each corpus. Our validation protocol comes from the one established in the scope of the TC-STAR project [10] and uses a randomly chosen 5% sample of the translated corpus for quality control. This protocol already defines a translation error typology as well as a way to measure the different problems encountered by the validators. Each of those errors is associated with a penalty point (see Table 1 for full details), the whole set of penalty points being used to compute a validation score.

Although the typology used is similar to that of the TC-STAR project, both the penalty points that are assigned to error types and the final validation score of the translations have been adapted to the needs of the Quaero project. Indeed, expectations in terms of quality are very high and our validation protocol aims at obtaining high quality translations. For that purpose, a threshold has been defined, which establishes that a corpus is rejected if the validation score is over 1 penalty point per 100 words.

However, validating the quality of translation is not a trivial task, not even for an expert, and each validation requires an analysis of the issues raised. Indeed, the different validations have given 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.

In the framework of our German-to-French production of corpus derived from transcriptions, the validation procedure has required two validations and thus one correction to reach the required quality level[6]. This takes into account neither the time spent handling the disagreements between the translation team and the validation team, nor carrying out the automatic format validations of the data resubmitted by the translation team.

For the current 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[7], one such corpus requires about 45-50 working days of production time. These are divided 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.

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). If this is the case, a new validation will be required.

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).

[6] The number of validations performed per corpus produced is rarely higher than three for most of our produced corpora.

[7] When these are not already part of the regular working team, professionals are tested in order to join the project.

5. Complexity and problems in the source corpus

Spontaneous speech is well known for showing a side of language structure which goes well beyond the scholarly learnt syntax. This may already look complex at a first glance, however, the day-to-day issues encountered by the translators go certainly much further.

It is due to the numerous discussions among our expert translators and proofreaders that we have been faced with the large complexity of the source data. In some occasions, the translation team has reported unexpected translation problems which were not initially covered by the translation guidelines. A number of these issues have required extensive discussion and evaluation with the Administration as well as consulting the end users with the aim of adopting the best solution according to their evaluation needs. Moreover, the work carried out by the translators and proofreaders is the result of a very close collaboration towards the production of a joint output. Translators have kept a follow-up of their discussions and decisions, which have been provided to the proofreaders. 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 that are spontaneous and grammatically fragmented by nature.

Furthermore, translation choices do not represent universal truths with a single possible solution. This is a general feature of the translation task, which has entailed a number of discussions with the validators and of disagreements with regard to ”wrong usage of target language” or to translation preferences. In fact, the concept of ”preference” needs to be taken into account during validation. Translating implies generating content in another language which most certainly offers a wide choice of possibilities. Limiting these 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. As translators, their work contains a creativity factor that is generally refrained when producing evaluation corpora and even the mere fact of having to translate from a source with mistakes, disfluencies or incoherences is often confusing.

The following sections illustrate those points which have been particularly difficult for the translators. Examples of translation are provided from German into French. Two categories are distinguished, depending on whether such points derive from phenomena specific to the speech domain or not.

5.1. Problems which are Specific to Speech Data

1. *Transcription segmentation* at the level of speech recognition system output represents a problem for the translation stage: the order of the elements within the target sentence can be different from that of the source sentence, and a source sentence may be divided into two or more segments where none of them is re-merged at a later stage. That way, when the segments of a source sentence must be translated into a language with a different grammatical structure but keeping the segmentation, the task may become very complicated as the translator needs to produce a good translation without deteriorating the alignment between the segments. This is the case of translation between French and German since the verbs are often placed at the end of the sentence in the latter.

   The explanation for having used such segmented data touches several aspects: data had been semi-automatically transcribed and segmented for the training and evaluation of speech recognition systems. Two key points in this were: cost and precision, bearing in mind that data needed to be well aligned to their audio. Thus, segments needed to be as small as possible, but not too small so as to make it reasonable cost-wise. Multiplying the number of frontiers would have increased the cost, thus, segment marquers were placed when long-enough silences or breathing pauses took place. So far this posed no problem, except that data came from broadcast news and very often speakers were communication experts who placed their silences and hesitations strategically in the middle of their sentences so as not to be interrupted and keep people’s attention. This is why we find segmentations in the middle of semantic units and why these broken sentences became an issue when moving on to the translation and alignment of the translations.

2. *The difficulty to understand transcribed data* has provoked a lot of discussions since translators have had to face either non understable source text or incomplete sentences, both of them regular phenomena from speech data. For certain translations, listening to the source data has been essential to allow translators to understand the transcription. Some translators’ choices may seem to contradict the guidelines, but we need to bear in mind that they had to choose alternative solutions, like in the following examples:

   - A source sentence contains ”die Landbrücke Mittel-und Zentralamerikas” but ”Mittelamerika” and ”Zentralamerika” refer to the same thing in French. The translator chose to translate this as ”l’isthme d’Amérique Centrale” (”Landbrücke Zentralamerikas”, i.e. *the isthmus of Central America*).
   - The translator did not understand what the speaker wanted to say with the term ”Ansätze” even after listening to the audio file. In fact, this is due to the speaker not articulating while talking and thus, making it impossible to understand

what he says. Several utterances from the same speaker present similar problems and, therefore, the translator had to interpret the transcription to proceed with translation.

3. *Transcription errors from the source audio data*, like spelling errors, missing words or misunderstood words and sentences, disturbed the translators, proofreaders and validators. When an important part of a source sentence is not understandable, this complicates the translator's task. We have observed that making the audio data available for the translators to use as reference is crucial. Whenever they were given access to them, translators managed to find the words to be translated and also to disambiguate problematic cases. Some examples follow below:

    - 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 intérieur 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.”.
    - One of the transcriptions talks about ”Mark Bartor” while it is actually the journalist ”Marc Bator”.

4. *The difficulty in understanding or interpreting the translation guidelines*, in particular when the translators need to deal with two different guideline points 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 ”-” symbols 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*).

    The translators did not always manage to follow the correct procotol when they faced a combination of those two guideline points (words partially pronounced and repeated), and thus, did not translate the repeated word but translated and tagged the partial word.

5.2. Other Encountered Problems

1. *The difficulty in establishing a balance between a translation that is close to the source text while rendering a fluent output in the target language.* This is always a source of disagreement among translators, proofreaders and, at the end of the process, validators, as it can be observed in the following examples:

    - 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.
    - At the time of the German Democratic Republic, ”deutschlandpolitische” referred to the issues concerning the political relationship between East and West Germany. However, since the translator is not meant to ”make explicit what is implicit” according to the guidelines, he has chosen a more literal translation with ”les questions de politique allemande” (i.e. *the issues in German policy*).

2. *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 desambiguate the transcription. Since this was not available, the translators remained as literal and close to the source sentence as possible, as shown in the coming examples:

    - 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.

- The segment "diesen Streit aus der Schuh des Manitu kennen zwölf Millionen Kinozuschauer mindestens." is hard to translate without any context. After having listened to the audio document, the translator learnt that the participants in a radio programme are listening to an excerpt from the movie "Schuh des Manitu", "Qui peut sauver le Far West", which has not been transcribed, but which is referred to by the speaker when saying "Douze millions de cinéphiles au moins connaissent cette dispute dans Qui peut sauver le Far West" (*At least twelve million movie-goers know this argument in Schuh des Manitu*).

6. Translation Issues and Improvements Carried out

For each corpus, when a set of sentences is rejected during a validation phase, these are resent to the translation team as guidance in the data correction they are asked to do (even if they are asked to correct the whole corpus). Corrections allow then to improve the translation and guarantee compliance with the specifications.

- A considerable number of errors indicated during validation have penalised what we call "wrong usage of the target language": even if they may correct from the orthographic and syntactic point of view, some sentences should not be rendered as such in the target language (they sound neither natural nor "native"). There have been many corrections and discussions concerning this matter, which as the reader may imagine, is a good source of disagreement.

- Translation fluency has also suffered, from time to time, from the lack of syntactic flexibility imposed by the need to stick to the initial segmentation and the fact of having to produce correct alignment between source and target text.

- Certain errors found in the source texts have been the cause of some deep consideration. In those cases, we have studied the corrections to be done directly with the translation teams so as to come up with appropriate solutions regarding the specifications and without wasting any unnecessary project time.

- The translation of titles for movies, TV series, broadcast programmes and books turned out to be an unexpected point of conflict. 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.

- Source data have been corrected, thus allowing the handling of format and encoding problems. A large number of encoding issues have been identified after the data has been processed by the translators with a variety of text editors, operating systems and translation tools that were used.

7. Conclusions

This article aims at giving an overview of the large complexity behind the production of parallel corpora for the evaluation of Speech MT systems, in the scope of the Quaero campaigns. The creation of reference translations from spontaneous speech transcriptions has revealed to be a challenge during the whole project. The interactions between the members of the production chain and the following of the detailed protocols for the translation, proofreading, validation and correction have proved decisive for the success of this work. ELDA has shown its expertise, which combined with the numerous exchanges and discussions with the project representatives (from both the IMMI and the DGA), have allowed to resolve the important number of encountered difficulties under very strict time constraints.

Although our data production has been carried out on a certain number of different translation directions, this study focused on a German-to-French parallel corpus produced from transcription data. These data have been particularly complex for translation experts, being used to our demands but forced to face numerous doubts and exceptions not covered by the project guidelines. Those difficulties have been classified as either specific to the speech data (such as issues coming from the audio segmentation, transcription errors, difficulties to understand fragmented audio segments, etc.) or to others. These latter cover issues related to either

(a) the production of translations that must be fluid but remaining close to the source data, or (b) the need for context, which is not always available in the source. We have detailed those cases and discussed the adopted approaches, as well as the improvements done following the different phases of quality control.

Last but not least, and still as part of the sustainability plan in data production, the parallel corpus described in this paper, together with the others mentioned, will be made available to the community once some remaining legal aspects are handled.

8. Acknowledgements

This work has been done and financed in the framework of the Quaero project. We would like to thank Hélène Mazo and Priscille Schneller for their work, as well as all the translators, proofreaders and validators for their efforts and interest in the project.

9. References

[1] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation," IBM Research Division, Thomas J. Watson Research Center, Tech. Rep., 2001.

[2] S. Banerjee and A. Lavie, "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*, Ann Arbor, Michigan, US, June 2005, pp. 65–72. [Online]. Available: http://www.aclweb.org/anthology/W/W05/W05-0909

[3] C. Callison-Burch, P. Koehn, C. Monz, K. Peterson, M. Przybocki, and O. F. Zaidan, "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*, Uppsala, Sweden, July 2010, pp. 10–17. [Online]. Available: http://www.aclweb.org/anthology/W10-17

[4] M. Eck and C. Hori, "Overview of the IWSLT 2005 Evaluation Campaign," in *International Workshop on Spoken Language Translation: Evaluation Campaign on Spoken Language Translation*, Pittsburgh, PA, USA, October 2005, p. 22.

[5] P. Koehn, "Europarl: A Parallel Corpus for Statistical Machine Translation," in *Conference Proceedings of the Tenth Machine Translation Summit*, Phuket, Thailand, 2005, pp. 79–86. [Online]. Available: http://mt-archive.info/MTS-2005-Koehn.pdf

[6] S. Strassel, C. Cieri, A. Cole, D. Dipersio, M. Liberman, M. Maamouri, and K. Maeda, "Integrated Linguistic Resources for Language Exploitation Technologies," in *Proceedings of LREC 2006*, 2006.

[7] NIST, "NIST 2009 Open MT Evaluation," NIST, Tech. Rep., 2009. [Online]. Available: http://www.nist.gov/speech/tests/mt/2009

[8] Z. Song, S. Strassel, G. Krug, and K. Maeda, "Enhanced Infrastructure for Creation and Collection of Translation Resources," in *Proceedings of LREC 2010*, Valletta, Malta, may 2010.

[9] M. Paul, M. Federico, and S. Stücker, "Overview of the IWSLT 2010 Evaluation Campaign," in *Proceedings of the seventh International Workshop on Spoken Language Translation (IWSLT)*, 2010, pp. 3–27.

[10] D. Mostefa, O. Hamon, N. Moreau, and K. Choukri, "Technological Showcase and End-to-End Evaluation Architecture, TC-STAR project," ELDA, Tech. Rep. Deliverable D30, May 2007.

[11] O. Hamon, A. Hartley, A. Popescu-Belis, and K. Choukri, "Assessing Human and Automated Quality Judgments in the French MT Evaluation Campaign CESTA," in *Proceedings of the MT Summit XI*, Copenhagen, Denmark, September 2007, pp. 231–238.

[12] O. Hamon and K. Choukri, "Evaluation methodology and results, MEDAR project," ELDA, Tech. Rep. Deliverable D5.3, November 2010.