# Diagnosing Human Judgments in MT Evaluation: an Example based on the Spanish Language

**Olivier Hamon**
ELDA
55-57 rue Brillat-Savarin
75013 Paris, France, and
LIPN, U. of Paris XIII
99av. J.-B. Clément
93430 Villetaneuse, France
hamon@elda.org

**Djamel Mostefa**
ELDA
55-57 rue Brillat-Savarin
75013 Paris, France
mostefa@elda.org

**Victoria Arranz**
ELDA
55-57 rue Brillat-Savarin
75013 Paris, France
arranz@elda.org

## Abstract

This paper aims at providing a methodology for analyzing the reliability of human evaluation in MT. In the scope of the second TC-STAR evaluation campaign, during which a human evaluation on English-to-Spanish was carried out, we first demonstrate the reliability of the evaluation. Then, we define several methods to detect judges who could bias the evaluation with judgments which are too strict, too permissive or simply incoherent.

## 1 Introduction

For a quarter of a century, many evaluation campaigns involving human evaluation in Machine Translation (MT) have been carried out and surely even more evaluations have taken place outside such campaigns. DARPA, and then NIST MT campaigns[1], among others, were certainly the most influential in human evaluation. However, recent evaluation campaigns such as IWSLT (Fordyce, 2007), TC-STAR (Mostefa et al., 2006), CESTA (Hamon et al., 2007) or WMT (Callison-Burch et al., 2007) have also highlighted the importance of human evaluation in MT. The results are checked carefully so as to assess system quality, especially due to the weakness of the automatic or semi-automatic metrics. However, what is not always highlighted are the inconsistencies of the human evaluation process, given that this remains the result of subjective judgments. It is particularly important to observe in detail how human judges react according to what they evaluate. Some measures have been defined to estimate a judge's consistency (Blanchon et al., 2004) or the number of judgments which are needed to have a relevant evaluation campaign (Koehn, 2007). It is well-known that inter-judge agreement is generally far from perfect (Ye and Abney, 2006), and even professional human translators disagree through different cases of translation. If this was not the case, one unique reference translation would be sufficient. However, how do judges evaluate a segment, depending on whether it is low or high quality? What are the difficulties met by judges which cause such lack of consistency among them?

Most of the previous evaluation campaigns have been carried out with English as a target language. However, some others have used languages with a richer morphology, such as Spanish or French. The answers we try to get in this experiment could help to improve the human evaluation set up, in particular when using morphologically richer languages like Spanish.

After describing the framework of our experiments we try to determine a methodology to find judges consistency and, if need be, to delete judges who would have done random evaluation. Finally, we draw some conclusions on our experiments.

## 2 Framework and General Results

The experiment presented here is done using the material from the TC-STAR second evaluation campaign (Mostefa et al., 2006). During this campaign, a human evaluation was carried out on

[1] http://www.nist.gov/speech/tests/mt/

English-to-Spanish direction, with data coming from European Parliament Plenary Sessions. The vocabulary used in these data belongs to the political and diplomatic domains.

The experiment involves three kinds of input: automatic transcriptions from Automatic Speech Recognition (ASR) systems, manual transcriptions (Verbatim) and Final Text Edition (FTE) data provided by the European Parliament. Each input has its own attributes and difficulties: the ASR input contains sentences with errors deriving from ASR systems; the sentences in the Verbatim input include spontaneous speech phenomena such as hesitations, corrections or false-starts; the FTE input sentences have been rewritten and do not include spontaneous speech phenomena.

Although we distinguish systems for ASR, Verbatim and FTE in the following results, we do not separate them, and thus we obtain a large range of scores, from the presumed lower quality ones (ASR) to the presumed better ones (FTE). 26 systems were evaluated within this human evaluation as a whole, which can be split up into 6 ASR systems, 9 Verbatim systems and 11 FTE systems. A subset of around 400 segments for each system output was used for the evaluation. Since each segment was evaluated twice, an overall of 20,360 segments were evaluated by 125 judges, corresponding to around 163 segments per judge. Judges were native Spanish speakers and did the evaluation through an interface available on Internet.

Each segment was evaluated in relation to both *adequacy* and *fluency* measures (White et al., 1994). For fluency, the quality of the language is evaluated and the judges had to answers to the question *"Is the text written in good Spanish?"*. A five-point scale was provided ranging from *"Spotless Spanish"* to *"Non understandable Spanish"*. For adequacy, automatic translations and corresponding reference segments were compared and the judges had to answer to the following question: *"How much of the meaning expressed in the reference translation is also expressed in the target translation?"*. A five-point scale was also provided to the judges ranging from *"All the meaning"* to *"Nothing in common"*. For both fluency and adequacy only extreme points were proposed on the scale, the rest of the points were unconstrained and then dependent on the judges' opinion.

The judges evaluated all their segments firstly according to fluency, and then according to adequacy. Thus, the fluency measure is applied independently and judges are not influenced by the reference translation. Both evaluations per segment are done by two different judges and no judge evaluates the same segment coming from two different systems. Finally, the segments are presented randomly.

The general results are shown in Figure 1.

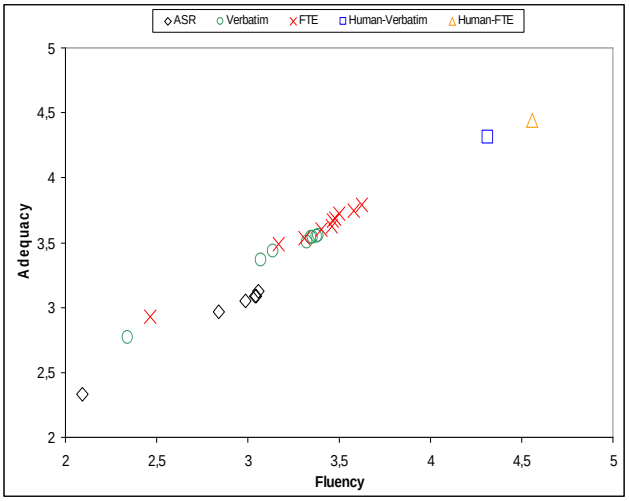

Figure 1: General results for fluency and adequacy.

Both Verbatim and FTE outputs located in the top right-hand side corner are coming from the human reference translations ("Human-Verbatim" and "Human-FTE", respectively) and are clearly higher than the automatic translations ("ASR", "Verbatim", "FTE"). Scores are better for FTE systems, then for Verbatim ones and, finally, ASR systems get the lowest results. This allows us to use a large set of sentence qualities and observe how judges evaluate accordingly.

## 3 Methodology and the Problem of the Human Evaluation

A main step when using human evaluation in MT is to define a protocol and a methodology to perform the test. Once the evaluation has been finalized by the judges, looking at the results is not sufficient. It is also important to know how reliable these judges are. Several methods can be used to determine the reliability of the evaluation, not giving the same information, but giving an indication about the performance of judges.

However, judgments are at any rate subjective. In this experiment, judges are not experts but end users and they react differently according to their condition, culture or knowledge. One of our goals is to determine how their judgments can be subjective. Then, we would like to define the kinds of segments that can pose a problem when reliability is low.

Then, the question we ask is: Are the Judgments "Correct"?

There are several ways to compute the agreement between judges. We present two of them here, a variation of the inter-judge agreement and the Kappa coefficient (Miller and Vanni, 2005). To go further, we try to detect whether some judges have particularly unfair results. This does not necessarily mean that judges are wrong, but that some of them could be too strict in comparison with the other judges.

### 3.1 Inter-judge agreement

Instead of computing a strict inter-judge agreement based on a binary agreement (two evaluators agree or disagree on a single segment), we have decided to measure an *n*-agreement, for which *n* is the upper difference between two scores of a same segment. For *N* segments, this is defined as follows:

$$n-agreement(n) = \frac{1}{N}\sum_{i=1}^{N}\delta(\left|S_i^a - S_i^b\right| \le n)$$

*where*:

$$\delta(\left|S_i^a - S_i^b\right| \le n) = \begin{cases} 1 \; if \;\; \left|S_i^a - S_i^b\right| \le n \\ 0 \; if \;\; \left|S_i^a - S_i^b\right| > n \end{cases}$$

*n*-agreement is described as the ratio of the number of segments for which the difference between the first evaluation of segment *S*, $S_i^a$, and its second evaluation, $S_i^b$, is lower or equal to *n*.

The results for the fluency and adequacy evaluations inter-judge agreement are presented in Table 1.

| Evaluation | Input | *n*-agreement | | | | |
|---|---|---|---|---|---|---|
| | | 0 | 1 | 2 | 3 | 4 |
| Fluency | FTE | .34 | .70 | .88 | .97 | 1 |
| | Verb. | .34 | .69 | .87 | .96 | 1 |
| | ASR | .29 | .63 | .85 | .95 | 1 |
| | Cumul. | .33 | .69 | .87 | .96 | 1 |
| Adequacy | FTE | .35 | .68 | .88 | .97 | 1 |
| | Verb. | .33 | .67 | .87 | .96 | 1 |
| | ASR | .30 | .66 | .84 | .95 | 1 |
| | Cumul. | .33 | .66 | .87 | .96 | 1 |

Table 1: Inter-judge *n*-agreement for the different types of data input.

Inter-judge agreement is quite similar whatever the data input or criteria of evaluation. Judges give exactly the same score for a third of the evaluated segments. This is quite low and demonstrates the relative subjectivity of the evaluation. However, around 70% of the evaluations do not differ in more than 1 point. Therefore, it seems more reasonable to use a 3 point scale instead of a 5 point scale.

We have observed that ASR input seems slightly harder to judge than Verbatim input, which is also slightly harder to judge than FTE input.

### 3.2 Calculation of the Kappa Coefficient

In addition to the inter-judge agreement we measure the global Kappa coefficient (Landis and Koch, 1977a), which allows to measure the agreement between *n* judges with *k* criteria of judgment. The measure goes further, taking into account the chance factor that judges give identical judgment on a same segment. For *N* segments, it is defined as:

$$\kappa = \frac{\overline{P_o} - \overline{P_e}}{1 - \overline{P_e}}$$

Where

$$\overline{P_o} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{n(n-1)}\sum_{j=1}^{k} n_{ij}(n_{ij}-1)$$

and

$$\overline{P_e} = \sum_{j=1}^{k}(\frac{1}{Nn}\sum_{i=1}^{N} n_{ij})^2$$

The amount of judges who evaluate the $i^{th}$ segment in the $j^{th}$ is represented by $n_{ij}$.

In other words, $\overline{P_o}$ is the proportion of observed agreement and $\overline{P_e}$ is the proportion of random agreement (also called *chance agreement*).

The values we obtain are shown in Table 2.

| Evaluation | $\overline{P_o}$ | $\overline{P_e}$ | $K$ |
|---|---|---|---|
| Fluency | .331 | .209 | .155 |
| Adequacy | .326 | .222 | .135 |

Table 2: Global Kappa coefficient values for fluency and adequacy.

According to (Landis and Koch, 1977a), $K$ values for both fluency and adequacy mean that judges agree slightly. But (Feinstein and Cicchetti, 1990) presented the limit of Kappa for low values even when agreement was high. This allows us to draw here some weakness of the Kappa coefficient at a practical level. It is representative of the exact comparison of the judgments, without taking into account the closeness of the judgments. One of its limitations is precisely that two judges who have close results would be penalized, as opposed to two judges with distinct results. This kind of case is particularly common in MT evaluation. Moreover, systematic errors between judges cause better coefficients since $\overline{P_e}$ would be lower.

One of the reasons for this low $K$ value can also be the number of judgments per segment, or the number of judges. But according to (Feinstein and Cicchetti, 1990), the minimal ratio is 6 evaluators for 30 segments, which seems impossible regarding our 10,380 segments for this experiment!

Finally, computing the Kappa coefficient does not provide better information about the inter-judge *n*-agreement, which informs more precisely about the reliability of the evaluation regarding different aspects of precision.

### 3.3 Methods for Detecting Outliers

When an evaluation is done, it is not easy to know whether judges do their evaluations seriously or not. This is particularly so if the judgments are not done by experts and with a large number of people. Judges can be more or less familiar with the tool they used to evaluate, some of them may be tired, or even not feeling well, etc. We should bear in mind that an overall evaluation can take around 2 or 3 hours, with or without pauses, which could cause a drop in the judge's attention.

To reduce the unavoidable subjectivity of the judgments, we try to locate outliers whose judgments are badly evaluated, if there are any. If these judges were detected, it may be useful to delete them from the evaluation set in order to homogenize the results and have a fair evaluation of systems. Three methods have been defined in order to detect those outliers.

**Mean score by Judge.** Each judge evaluates a subset of around 163 segments. This subset has been built randomly and should be representative of the whole set of segments (10,180 segments). Since each segment has been evaluated twice, we can compute the mean score of the judge on his subset and compare it with the score of the same subset obtained with other judges.

Figure 2 and Figure 3 show the mean score for fluency and for adequacy, respectively. Judges are ranked increasingly, so as to have judges' scores sorted from the lowest to the highest.

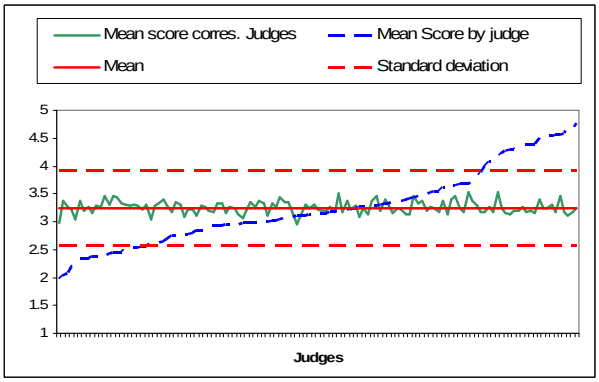


Figure 2: Mean score by judge for fluency and mean score for corresponding judgments from other judges.

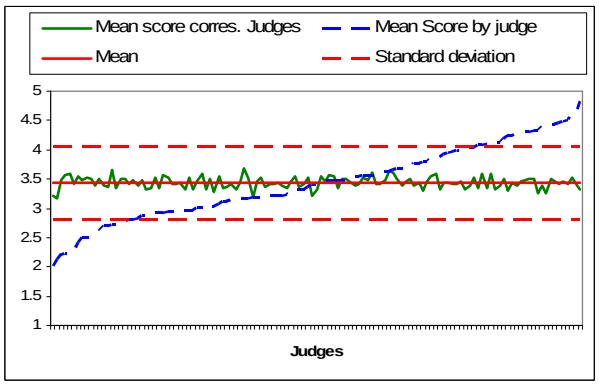


Figure 3: Mean score by judge for adequacy and mean score for corresponding judgments from other judges.

The variation of mean score by judge is similar for both fluency and adequacy. As expected, the score of each subset (plain peaky curve, Figures 2 and 3) is close to the general score of the whole set of segments (plain straight line). So each judge's subset is a representative sample of the whole data.

What is more surprising is the curve of the mean score by judge (dashed curved lines, Figures 2 and 3). We can see that some judges gave very low or very high scores compared to the other judgments on the same subset of segments.

We suspect that these evaluators misunderstood the 5-point scale or did not pay enough attention to the evaluation, or are either too strict or not strict enough. Judges above and behind the standard deviation are deviant and could probably be turned down to homogenize the judgment or be asked to redo their evaluation and thus obtain a more objective evaluation.

This method allows us to compare the score of each judge with the score of his subset of segments. But of course, for a given judge, we can have a mean score that is very close to the mean score of his subset with big differences for each segment. This is why we investigated the mean agreement by judge.

**Mean agreement by judge.** For each judge, a distance score is computed between his own judgment on a segment and the corresponding judgments from the other judges on the same segment. In other words, a mean agreement is measured for each judge comparing his own judgments to those of the other judges who assessed the same segments.

Figure 4 and Figure 5 present the mean agreement for fluency and for adequacy, respectively. Once again, judges's scores are ranked in an increasing manner.

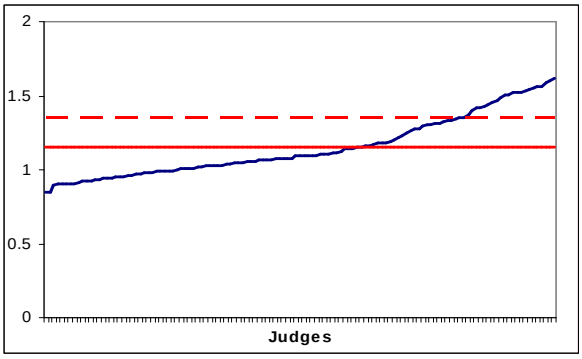


Figure 4: Mean agreement for fluency.

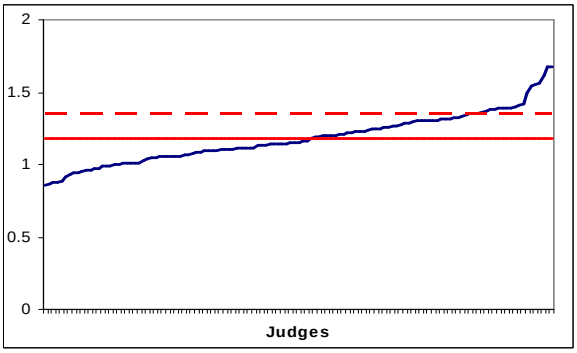


Figure 5: Mean agreement for adequacy.

As for the *Mean Score*, fluency and adequacy curves follow the same trend, although the adequacy one increases faster when agreement is higher. For certain judges, mean agreement is above 1.5, which is quite high since the largest mean agreement possible is 4. This means that these judges disagree with other judges in 1.5 in mean. It does not necessarily mean that these judges have not done their judgments correctly (what is even more, they are close to the other judges), but simply that some judges are stricter than others.

**Easy sentences.** Some sentences are easier to translate than others, because of their length, their simpler lexical or syntactic content, etc. We decided to make a selection of those "easy sentences" in order to observe the judgments done. In theory, those sentences should not represent any problem for automatic systems: these systems should not make any mistakes, and then judgments should be "perfect". Thus, a lower judgment draws our attention to the judge who has done it if the automatic system actually managed to translate the sentence correctly. Easy sentences can be described as containing few words, or easy words to translate like "gracias". They can also be sentences which occur frequently in the data (even in the development data to train the systems). Figure 6 illustrates some of those sentences.

¿ podría hacer más la Comisión ?
gracias , Presidente .
la respuesta es compleja .
gracias .

Figure 6: Examples of "easy sentences".

A total of 80 segments have been manually identified, which should allow us to identify evaluators for whom evaluations are incoherent.

The study of the translated sentences and their judgments shows some segments that are not correctly assessed, which does not mean that the general results of the judge who assessed them are not correct either. Such segments are localized and reasons for such erroneous assessment could be fatigue, lack of attention, or others things which are more linked to the activity than to the judge's competence itself.

For detecting incorrect segments, a study should be done at a segment level. But currently it is hard to provide such a study since there are only two evaluations per segment and, more particularly, because of the tedious and time-consuming work to be done. Moreover, the proportion of such segments seems to be very low and in any case, these segments are drowned in the whole volume of segments.

However, even if our analysis is quite subjective, some judges seem to evaluate incorrectly a significant amount of easy sentences.

### 3.4 Removing Outliers

The standard deviation allows to observe the statistical dispersion of judges away from the mean. Thus, we can remove judges who are above the positive standard deviation (for mean score and mean agreement) and under the negative standard deviation (only for mean score). Then, Table 3 can be drawn to compare the judges deleted for each method.

| | Mean Score | Mean Agreement | Easy Sentences |
|---|---|---|---|
| Fluency | 45 | 23 | 9 |
| Adequacy | 45 | 18 | 10 |
| Fluency + Adequacy | 30 | 10 | 4 |

Table 3: Number of judges deleted with the three methods.

Moreover, for fluency, 20 judges are common to both *Mean Score* and *Mean Agreement*, while for adequacy there are only 15. Most of them are included in the upper part of the *Mean Score* (17 for fluency, 5 for adequacy), the others in the lower part (3 for fluency, 10 for adequacy). It seems that outliers are too permissive for fluency, but on the contrary they are too strict for adequacy. Indeed, for fluency, judges have only the translated segment to evaluate, they have nothing to compare with and then are more flexible regarding the different possibilities of judgments. However, when comparing to the reference segment for adequacy, judges are then able to try to match exactly both segments. Another possible reason for being more permissive regarding fluency could be that an MT user's expectations are always higher with regard to content transmission than with regard to syntactic perfection, or rather, that a system's user will mind less having a syntactically imperfect output than a semantically inaccurate one.

Should we decide to delete those judges, that means that more than a third of the judgments would be deleted for *Mean Score*, that the number of judgments for *Mean Agreement* would be divided by 6, and finally divided by 13 for *Easy Sentences*.

This experiment may not mean that we delete "bad judges", but rather that we only delete judges who diverge from the set of judges. Thus we homogenize the evaluation.

After deleting judges and their judgments, we have computed again the scores of the human evaluation. Table 4 shows the Pearson correlations between the scores of the official evaluation presented above, and the scores after deleting judges, for the three methods of identification.

| | Mean Score | Mean Agreement | Easy Sentences |
|---|---|---|---|
| Fluency | .98 | .99 | .98 |
| Adequacy | .99 | 1.00 | .99 |

Table 4: Pearson correlations between official scores and scores after deleting judges.

Spearman's rank correlation is up to .99 for all the methods and criteria.

The results are not really surprising for the *Mean score*: judges who have higher and lower mean scores have been deleted and they about complement each other.

However, this is more surprising for the *Mean agreement*. Deleted judgments are in strong disagreement with the judgments from other judges, so scores should be from the boundaries and they bias strongly the results. In fact,

comparing “good judges” with outliers, scores are not identical but very close: For fluency, mean scores are 3.47 against 3.24 and 3.26 against 3.91 for adequacy mean scores, respectively. Values for deleted judges are still low regarding other judges, and they are still representative for the whole evaluation set.

Regarding *Easy Sentences*, the amount of judges removed is probably not sufficient to affect the scores, all the more so, as according to the previous comments, there are no real divergent judgments.

Although general results are higher, or lower, the trend of results is identical, like the systems ranking, as shown in Figure 7.

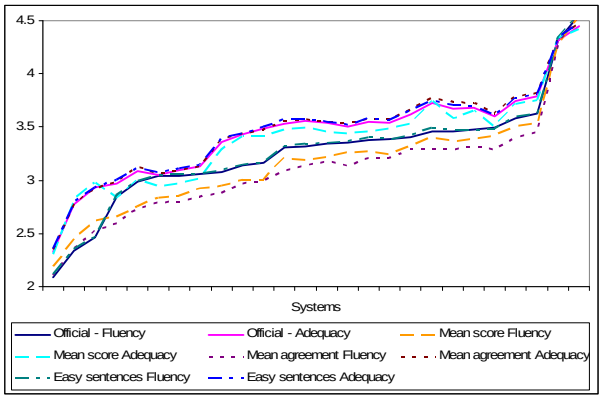


Figure 7: System scoring for Official results and the methods of judges deletion.

Another interesting point would be to know whether low agreement means wrong scores. In the same way as above, we have computed the Pearson correlation between official results and scores from the deleted judges only. Pearson correlation for fluency is .84, and for adequacy .96. This does not make a strong difference between the results for adequacy. However, this difference is more important for fluency. As mentioned earlier, that reflects bigger differences between judges for fluency (but not particularly higher difficulties for evaluating it), because of the absence of comparison to a reference. Since judges are typically free in their evaluation (i.e. there is no detailed guideline), they are more heterogeneous than for the adequacy evaluation, during which they refer to a single reference (and, which, in a certain way, serves the purpose of a guideline).

However, keeping all the judges does not really affect the systems ranking either for fluency or for adequacy, although, in general, scores are slightly lower.

## 4 Conclusion and Further Work

Our experiment is based on three general points in order to diagnose the performance of human judges in machine translation in two ways: a statistical observation of the judgment, and a linguistic study of the evaluated sentences.

First, we observed the agreement between judges to estimate the reliability of the evaluation. We drew the conclusion with the inter-judge *n*-agreement that this experiment contains an extremely detailed scale of judgments (5 points), which seems to confuse the evaluators, and we propose to limit the criteria to three. It would be interesting to make the same observations taking into account three criteria, for instance by merging criteria “1” and “2” , and “4” and “5”, and then studying the difference. Using the Kappa coefficient has proved its limitations in a practical case, since it does not take into account the variation of the agreement.

Then we tried to define a protocol and methods for detecting outliers, i.e., judges who are too subjective regarding other judges. In that experiment, deleting this kind of judges did not clearly change scores and ranking. Moreover, the number of judgments done does not allow to change the score so easily when deleting several judges.

Our future work will consist in applying this method to the third evaluation campaign of the TC-STAR project (under the same conditions but different judges), and to French corpora from the CESTA evaluation campaigns. This should allow us to observe the consistency of judges and perform intra-judge agreement too.

We also would like to do the same kind of study, but this time according to segments, systems and data criteria. Although it is important to measure the reliability of human evaluation, we also need to find how to improve the methodology and, most of all, to understand why judges evaluate sentences in such a way. This is directly linked to a currently-ongoing linguistic study of the evaluated segments and how these reflect the judges’ criteria and skills.

## Acknowledgement

This work was supported by the TC-STAR project (grant number IST-506738).

## References

Blanchon H., Boitet C., Brunet-Manquat F., Tomokio M., Hamon A., Hung V. T. and Bey Y. 2004. *Towards Fairer Evaluation of Commercial MT Systems on Basic Travel Expressions Corpora.* Proc. IWSLT 2004. Kyoto, Japan.

Callison-Burch C., Fordyce C. and Koehn P., Monz C. and Schroeder J. 2007. (Meta-) Evaluation of Machine Translation Proceedings of the Second Workshop on Statistical Machine Translation, June 2007, Prague, Czech Republic.

Feinstein A.R., Cicchetti D.V. 1990. *High agreement but low kappa: I. The problems of Two Paradoxes.* J. Clin. Epidemiol., 43, 543-548.

Fordyce C. 2007. Overview of the IWSLT 2007 *campaign.* In Proceedings of IWSLT 2007, Trento, Italy.

Hamon O., Hartley A., Popescu-Belis A. and 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. 2007. *Evaluating Evaluation Lessons from the WMT 2007 Shared Task.* MT Summit XI Workshop on Automatic Procedures in Machine Translation Evaluation, September 2007, Copenhagen, Denmark.

Landis J.R., Koch G.G. 1977a. *The Measurement of Observer Agreement for Categorical Data.* In Biometrics, 33, 159-174.

Landis J.R., Koch G.G. 1977b. *A one-way components of variance model for categorical data.* Biometrics, 33, 671-679.

Miller K.J., Vanni M. 2005. Inter-rater Agreement *Measures and the Refinement of Metrics in the PLATO MT Evaluation Paradigm.* In Proceedings of MT Summit X, Phuket, Thailand.

Mostefa D., Hamon O. and Choukri K. 2006. *Evaluation of Automatic Speech Recognition and Spoken Language Translation within TC-STAR: results from the first evaluation campaign*, Proc. Language Resources and Evaluation Conference, Genoa, Italy.

Mostefa D., Garcia M-N., Hamon O. and Moreau N. 2006. *Evaluation report, Technology and Corpora for Speech to Speech Translation (TC-STAR) project.* Deliverable D16.

White J. S. and O'Connell T. A. 1994. *The ARPA MT Evaluation Methodologies: Evolution, Lessons, and Future Approaches.* Proceedings of AMTA Conference, 5-8 October 1994, Columbia, MD, USA.

Ye Y. and Abney S. 2006. *How and Where do People Fail with Time: Temporal Reference Mapping Annotation by Chinese and English Bilinguals.* In Proceedings of Frontiers in Linguistically Annotated Corpora 2006, a Merged Workshop with 7th International Workshop on Linguistically Interpreted Corpora (LINC-2006) and Frontiers in Corpus Annotation III at ACL, Sydney, Australia, pp13-20.