Is my Judge a good One?

Olivier Hamon

ELDA, 55-57 rue Brillat-Savarin - 75013 Paris, France
LIPN, Université de Paris XIII, 99 avenue J.-B. Clément - 93430 Villetaneuse, France
hamon@elda.org

Abstract

This paper aims at measuring the reliability of judges in MT evaluation. The scope is two evaluation campaigns from the CESTA project, during which human evaluations were carried out on fluency and adequacy criteria for English-to-French documents. Our objectives were threefold: observe both inter- and intra-judge agreements, and then study the influence of the evaluation design especially implemented for the need of the campaigns. Indeed, a web interface was especially developed to help with the human judgments and store the results, but some design changes were made between the first and the second campaign. Considering the low agreements observed, the judges' behaviour has been analysed in that specific context. We also asked several judges to repeat their own evaluations a few times after the first judgments done during the official evaluation campaigns. Even if judges did not seem to agree fully at first sight, a less strict comparison led to a strong agreement. Furthermore, the evolution of the design during the project seemed to have been a source for the difficulties that judges encountered to keep the same interpretation of quality.

1. Introduction

Many evaluation campaigns are carried out in machine translation using human judgments. The existing literature has defined a number of human evaluation criteria: intelligibility, fidelity, reading time, correction time, etc. (van Slype, 1979). Currently, generally used evaluation criteria are *fluency*, *adequacy*, or, less used, *informativeness* (Carroll, 1966; White et al., 1994). Lately, (Eck and Hori, 2005) tried to extend the adequacy scoring with the *meaning maintenance* criteria. And more recently, the 2008 NIST campaign introduced *preference* criteria allowing comparison between systems[1].

However, literature contains rather few methods to check the reliability of judges' behaviour, and what is more, the fluctuation of their performance regarding their environment. A study of consistency has been made in (Blanchon et al., 2004), resulting in the definition of some measures to test the reliability of judges, as well as the number of human judgments needed (Koehn, 2007). Whatever it is that affects annotations or judgments, one can denote that most of the time humans disagree (Ye and Abney, 2006). Here, we reuse a methodology already described in (Hamon et al., 2008) that was followed on Spanish data in order to test the reliability of MT judges. In this paper, we will focus our study on French data from the CESTA project (Hamon et al., 2007) and extend the study: after analysing the inter-judge agreement, we will describe a study of intra-judge agreement on the same data, but with a time delay between the first and the second evaluation. Finally, we will also study the impact of the 'judgment design' on the judges' behaviour.

2. Framework of the Experiments

We based our experiments on the data of two French MT evaluation campaigns organised within the CESTA project (Hamon et al., 2007), particularly on the English-to-French direction task (a second task on the Arabic-to-French direction was also carried out). For MT systems, the objective of the first campaign was to translate into French 790 English segments (i.e. sentences most of the time) from the Official Journal of the European Community (JOC) corresponding to around 20,000 words. Five systems participated in this task. The objective of the second campaign was to translate 917 segments from the ”Health Canada” Web site[2], i.e. from the health domain, also corresponding to around 20,000 words. Five systems participated in this second task and not all of them were the same as for the first campaign. For each campaign, four reference translations were available for the need of an automatic evaluation. A reference translation was evaluated during the second evaluation campaign, as opposed to the first one for which only MT systems were evaluated.

Human evaluations were included into the evaluation process of the two campaigns. Each segment was evaluated according to *fluency* and *adequacy* criteria by two different judges. The segments were presented randomly to the judge and a maximum of 90 minutes was arbitrarily fixed per process. This corresponds to around 150 judgments (depending on judges). Indeed, it seems that beyond this limit judges lose their focus.

For *fluency*, the judges were asked to answer the question ”Is this text written in good French?” for each segment by giving a score on a 5-point scale, which went from ”Native French” (5) to ”Non understandable” (1). For *adequacy*, they were asked to compare the meaning of the evaluated segment to that of a reference translation and give a score on a 5-point scale, going from ”The whole meaning is present” (5) to ”Nothing in common” (1).

For the first evaluation campaign, 112 judges evaluated the 3,950 segments from the English-to-French task, i.e. around 71 segments per judge (in addition to these segments, other segments from the Arabic-to-French task were also evaluated in the meantime). For the second evaluation

[1] http://www.nist.gov/speech/tests/mt/2008/doc/

[2] http://www.hc-sc.gc.ca

campaign, only a subset of all the segments was evaluated, corresponding to 3,456 segments evaluated by 48 judges (again, for the English-to-French task), i.e. around 72 segments per judge. The number of evaluations done per judge were then similar from one campaign to the other.

A web interface was especially developed to help the human judgments and store the results, but some design changes were made between the first and the second campaign. Intermediate values of the 5-point scales were explicitly named for the first campaign, while the judges were free to define their own intermediate values for the second campaign. So, only boundary values were named in the second case. The second distinction was about the order of fluency and adequacy evaluations: both evaluations were done in parallel for the first campaign (each segment was first evaluated according to fluency and, then, to adequacy), while evaluations were separated for the second campaign (first, all segments were evaluated according to the fluency criterion, then all these segments were evaluated according to the adequacy criterion). Those changes are studied in more detail in the following sections.

3. Reliability between Judges

As previously said, it is essential to analyse the inter-judge agreement, first because of the subjectivity of the judgments and the help it provides for the analysis of evaluations, but also because it may allow to detect "problematic" judges (as well as judgments), outliers, and potentially eliminate them. To that aim, we have used the methodology presented in (Hamon et al., 2008), as a variation of the inter-judge agreement and Kappa coefficients (Miller and Vanni, 2005) in order to detect the outliers (in particular using the mean agreement per judge).

3.1. Inter-judge n-Agreement

First, we compute the inter-judge *n*-agreement, for which *n* is the upper difference between two scores of a same segment. For *N* segments, the *n*-agreement is stated as follows:

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

where:

$$\delta(|S_i^a - S_i^b| \leq n) = \begin{cases} 1 & \text{if } |S_i^a - S_i^b| \leq n \\ 0 & \text{if } |S_i^a - S_i^b| > n \end{cases}$$

The *n*-agreement is defined 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 than or equal to *n*.

The results for the fluency and adequacy evaluations inter-judge *n*-agreement are shown in Table 1, for the first campaign (Run 1) and the second one (Run 2).

For close to 40% of the segments, judges give identical scores (e.g. 1 for a judge A and 1 for a judge B, *n*=0). This low percentage shows a consequent subjectivity in the judgments given by the judges. Althouh we could assume that fluency is easier to evaluate than adequacy, the inter-judge agreements are not particularly higher for one or the other in either campaign. Likewise, it is difficult to say that the inter-judge agreements are higher for a domain or another, when comparing the two campaigns, since the values do not show a general trend. This last point is quite surprising, since we may also assume that using a particular vocabulary (here, from the health domain) makes evaluation harder.

Evaluation		*n*				
		0	1	2	3	4
Run 1	Fluency	.39	.84	.97	1	1
	Adequacy	.36	.76	.93	.99	1
Run 2	Fluency	.41	.78	.94	.99	1
	Adequacy	.47	.80	.93	.99	1

Table 1: Inter-judge *n*-agreement [0-1] for the two evaluation campaigns.

As usual in that kind of experiment, inter-judge *n*-agreement, when $n > 0$, shows a good agreement among judges, meaning that evaluation is nevertheless reliable and judgments are quite stable. The trend is similar for the two campaigns, event if data are from different domains and vocabularies. Although agreements are not perfect, we are convinced that it is not possible to reach a 100% of at least *1*-agreement, since judges do not have the same perception of what a "good" translation is. In our experiments, we took the point of view of potential users, and we decided not to provide them with much information that could guide them on purpose: that was also one of the reasons why the web interface used to collect the judgments was changed for the second evaluation campaign.

3.2. Global Kappa Coefficient

Next, we compute the Global Kappa coefficient (Landis and Koch, 1977), which allows to measure the agreement between *n* judges with *k* criteria of judgment (here, *k* corresponds to the 5 values of the scales), taking into account the chance factor that judges give identical judgment on a same segment. For *N* judgments, it is stated 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} criterion 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 (i.e. *chance agreement*). The values of kappa coefficients for the two campaigns are shown in Table 2.

Following previous experiments, agreements according to Kappa coefficients are low, not to say very low. That is an issue already raised in (Feinstein and Cicchetti, 1990) and it

Evaluation		$\overline{P_o}$	$\overline{P_e}$	κ
Run 1	Fluency	.394	.233	.210
	Adequacy	.364	.215	.189
Run 2	Fluency	.516	.267	.340
	Adequacy	.566	.280	.398

Table 2: Kappa coefficient values [0-1] for the two evaluation campaigns.

confirms, for different data, that the Kappa coefficient does not provide more information about judges reliability than *n*-agreement.

3.3. Outliers Detection

The results of the agreement between the judges make us wonder whether judges are really reliable or not. It may be that some judges are mistaken, do not understand the evaluation task or, what is even worse, do not pay attention to their judgments. The duration of the evaluation may be long and fatigue can affect the precision of judges. A methodology to detect questionable judges has already been proposed in (Hamon et al., 2008) and we apply here one of the methods for computing the mean disagreement per judge.

This method consists in computing, for each judge, a distance score between his own judgment on a segment and the corresponding judgment from the other judge on the same segment, and likewise for all his segments (the other judges are then not always identical from one segment to another). In other words, the objective is to rank judges according to their strictness, and detect potential judges who would be ”outliers”, i.e. judges who would not do their evaluations correctly.

Figures 1 and 2 show the results of the disagreement per judge for the first campaign and the second campaign, respectively. Judges are ranked according to increasing scores.

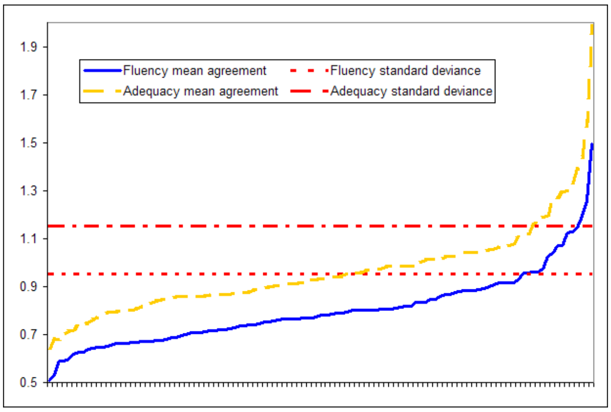

Figure 1: Fluency and adequacy mean disagreements for the first campaign. The list of judges is presented in X-axis.

Mean disagreements per judge are quite high, as we could find in the state-of-the-art, higher than 0.5 and might be above 1.9 for certain judges. For the first evaluation campaign, the means (of the mean disagreements per judge) are of 0.80 for fluency and 0.96 for adequacy, meaning judges disagree with other judges in 0.80 in mean (on a maximum of 4). For the second evaluation campaign, they are of 1.43 for fluency and 1.40 for adequacy. For both fluency and adequacy evaluations, second campaign disagreements between judges are higher, as well as the standard deviances, but curves are straighter. This means judges' behaviour is quite the same during the second campaign, contrary to the first campaign for which only several judges clearly move away from the trend. This leads up to detect these judges, the outliers, in observing the dispersion of judges. In order to establish a boundary, we take judges above the standard deviance. Table 3 shows the number of judges detected for each campaign and each criterion.

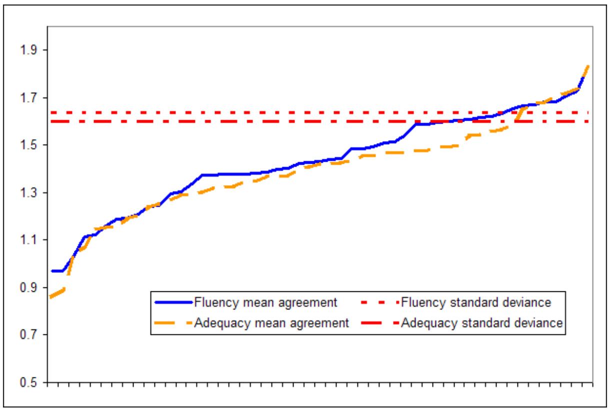

Figure 2: Fluency and adequacy mean disagreements for the second campaign. The list of judges is presented in X-axis.

Evaluation	Fluency	Adequacy	Common
Run 1	15	13	6
Run 2	8	7	4

Table 3: Number of detected judges per campaign and criterion.

Around 15% of judges can be considered as outliers. It means they are just divergent, they are not ”bad” judges: as previously said, their results may be due to various reasons and not only to their competence.

However, even if we delete these outliers, system scores are very similar and correlations between the official scores and scores after the deletion of judges are close to 100%. So, this allows us to determine that the evaluation is reliable, even if there are several less coherent judges. The amount of judges and the mean of scores are sufficient to have a valid evaluation and therefore, those 15% outliers do not have a strong impact on the final results.

4. Intra-judge Agreement and Impact of the judgment design

After having studied the inter-judge agreements to test the reliability of an evaluation, we know some of the potential difficulties met with judges. However, the results are not objective enough to know if our judges are ”good ones”: the inter-judge agreement allows to consider whether judges perceive the quality of a translation in the same way, not whether one specific judge is making correct judgments. Therefore, intra-judge agreement allows to consider the reliability of one judge alone, and observe if his judgments are consistant.

We proceed in a slightly different way from that in (Callison-Burch et al., 2007), where the authors took 10% of the overall set of a judge to be evaluated twice by the same judge, within the same evaluation session. However, we believe that the delay between two evaluations of a same segment is not sufficient, as the judge may still have his judgment in mind, unconsciously or not. So, we asked several judges to repeat their own evaluations a few times after the first judgments done during the official evaluation campaigns: delays between four and twelve months separated the two evaluations. Unfortunately, only three judges participated in that experiment, out of the four judges that participated in both evaluations, and due to time and cost constraints it has not been easy to convince people. Still, this has allowed us to study the results they obtained in more detail.

4.1. Representativeness

A first step in that experiment is to observe if the three judges (namely ”Judge A”, ”Judge B” and ”Judge C”) are representative of all the judges of the two campaigns, and if their scores correlate with the overall system scores (as well as their ranks).

Judge B is the only judge to be an outlier for each evaluation (except for the fluency evaluation of the first campaign). But as previously said, making an outlier of a judge does not mean he/she is a ”bad” judge, not to say his/her results are representative of the overall results. This is confirmed by the correlation scores between the overall results and the individual judges' results. Indeed, comparing the results per judge within the official results of the campaigns, mean of the overall judges, only Judge B gets a correlation score of 0.60 for the fluency evaluation of the first campaign (which is surprisingly the only one for which the judge is not an outlier). Other correlation scores are above 0.80, generally above 0.90. So, we consider the three judges as representative, but taking the results ”with a pinch of salt” nevertheless.

4.2. Methodology

The objective of our experiment is twofold: first, we want to observe the intra-judge agreement, but we also try to study the impact of the judgment design on the judgments. Having that aim in mind, two designs were available: the one used for the first evaluation campaign and the one used for the second evaluation campaign. The latter differs in two factors: fluency and adequacy evaluations are done separately, and descriptions of the 5-point scales are only explicit for the highest and lowest values. We then cross-checked the evaluations in order to observe the differences of judgments, according to the following protocol:

- Judges do an initial evaluation of their sets of segments during the first campaign, with the first design (i.e. coming from the official campaign);
- Judges do an initial evaluation of their sets of segments during the second campaign, with the second design (i.e. coming from the official campaign);
- A delay of a few months takes place during which no evaluation is done;
- Judges evaluate a second time their sets of segments from the first campaign, with the second design;
- Judges evaluate a second time their sets of segments from the second campaign, with the second design;

Thus, only the re-evaluation of the second campaign set allows us to compute a real intra-annotator agreement, and notice the differences in judgment for a same judge and for a given time. In addition, the re-evaluation of the first campaign set allows us to observe the impact of the judgment design change, by comparing with the re-evaluation of the second campaign.

4.3. Pearson correlations

We first compute the Pearson correlations between the first and second evaluations at the level of system scores. This is shown for the three judges, the two evaluation criteria and the two campaigns according to the scoring (Table 4) and the ranking (Table 5) of MT systems.

Judges	Run 1		Run 2	
	Flu.	Ade.	Flu.	Ade.
A	.47	.63	.96	.99
B	.96	.66	.97	.92
C	.98	.62	.90	.89

Table 4: Pearson correlations [0-1] for the two campaigns, according to the scoring.

Judges	Run 1		Run 2	
	Flu.	Ade.	Flu.	Ade.
A	.60	.60	.92	1
B	.90	.50	.95	.98
C	1	.50	.92	.89

Table 5: Pearson correlations [0-1] for the two campaigns, according to the ranking.

Then, to observe the consitency of judges, we focuses on the Pearson correlation results of the second campaign, since the scores are comparable as opposed to the results of the first campaign, for which the judgment design is not the same. Correlations are high, but not as high as we may imagine, going from 0.90 to 0.99. That means judgments variations are present in judges. Moreover, there is no special trend according to fluency or adequacy, it rather depends on the judges.

However, correlations do not reflect another reality, when scores are lower or higher in the repeated evaluation. The absolute value of difference means on system scores goes from 0.08 to 0.42 (on a scale from 1 to 5) for both fluency and adequacy, which represents a consequent variation (most of the differences are statisticaly significant, but not all of them). The highest difference we found on one unique system was of 1.08 but the general trend is around 0.4. We should also notice scores are generally lower, especially regarding the adequacy evaluation. The ”experience” of judges probably plays an important role here, and that may mean judges do not stay long enough on segments to analyse them. Actually, even after a few months, some judges told us they remembered having seen several of these sentences, without remembering which values they had assigned to them.

Regarding the correlation on ranks, they are most of the times just under 1, which is very high. The fact of being slightly under 1 is generally due to the inversion of two systems within the ranking (or due to the changing of equally-ranked systems) when one is close to the other. Furthermore, ranking results show a particular trend of agreement: when a system is above another system it will most probably remain in that position in another evaluation too, even if scores vary and are higher or lower.

4.4. Intra-judge *n*-agreement

To go further, we compute an intra-judge *n*-agreement for each judge, as described above for the inter-judge *n*-agreement. Results are shown in Table 6.

Evaluation		*n*			
		0	1	2	3
Fluency	Judge A	.44	.88	.99	1
	Judge B	.44	.74	.86	.96
	Judge C	.47	.86	1	1
Adequacy	Judge A	.63	.93	1	1
	Judge B	.43	.74	.88	1
	Judge C	.56	.97	1	1

Table 6: Intra-judge *n*-agreement [0-1] for the second evaluation campaign.

With the exception of the *0*-agreement, results are very high, more than for the inter-judge agreement if we compare both. However around half of the segments do not have the same judgments when the evaluation is repeated. Moreover, an important number of segments shows differences above 2 points. It seems more difficult to find here the judgments for fluency than for adequacy. This could be explained by the absence of referential for fluency and thus the difficulty to compare the quality of segments without a gold standard.

4.5. Impact of the judgment design

An interesting point of our experiment concerns the study of the impact of the judgment design. Indeed, the context of an evaluation is essential when it concerns humans. A number of factors should be considered for a same judge (without comparing two judges according to their knowledge, culture, etc.): fatigue, environment, interruptions, being disturbed, interface, etc. We focus here on the design aspect, since two factors were modified on the web interface: the order to evaluate fluency and adequacy (either together or separately), and the display of 5-point scales, with or without the intermediate points explicitly defined.

We are aware of the fact that changing two factors at the same time is difficult to analyse, but we assume that:

- Changing the order of evaluation should only affect the adequacy scoring. When the two criteria are evaluated together, one segment evaluation is longer and the judge already gives an estimation of quality, according to fluency, which could influence his judgments for adequacy. Moreover, his attention goes from one criterion to another, while in the second case it only focuses on one criterion at a time. Since fluency is presented first for a new segment, only the adequacy score should be different.

- Changing the definition of scales, by deleting the definition of intermediate points, should affect both fluency and adequacy, since both criteria are modified at the same time.

Looking at Tables 4 and 5, the results allow us to compare a repeated evaluation with a modified design to a repeated evaluation with a non-modified design. According to the results of the first campaign (Run 1), adequacy correlations are low for all the judges, while only one judge gets low correlation for fluency. At the same time, all the correlations are very high for the second campaign (Run 2), which could be considered as the gold standard of our experiment. At first sight, the modification of the design makes the interpretation of fluency and adequacy different to judges. The assumption stated above shows the three evaluators were disturbed by changes in the order to present fluency and adequacy in the first campaign, and their judgments are strongly modified accordingly. Next, only one judge presents some differences with fluency evaluation in the first campaign, showing the changes of the 5-point scale description could also disturb judges, probably at a lesser degree.

Actually, we can not tell which judgment design is the best, even if we prefer the design separating fluency and adequacy judgments according to the comments stated above. Yet, design changes are most certainly affecting the scores and then the entire evaluation, although it should also be mentioned that this experiment should be done on a more representative sample of judges to study the impact of the design in more detail.

5. Conclusions and further work

This paper describes an experiment to estimate the reliability of judges who carry out evaluation on MT systems. First a methodology to measure the inter-judge agreement has been re-used on English-to-French data. This shows that judges do not seem to agree fully at first sight, but obtain strong agreement when the comparison is not so strict. Then, we tried to detect judges who seem to fail in the evaluation so as to study what happened.

Next, we estimated the reliability of judges alone, by looking at their intra-agreement, i.e. how they agree with their own evaluations. Results are quite surprising, since judgments of a first evaluation compared with a repeated evaluation show strong variations. Last but not least, we observe in this experiment that judgment design may influence the judges' behaviour in several ways. Each design parameters has its importance, and we should put into perspective scores acquired from judgments according to the scope they are obtained in.

In order to study further the problems detected and described above, further experiments should be carried out on the segments evaluated. This should hopefully help us detect problematic segments and difficulties met by the judges, as well as to confirm or invalidate our observations. Indeed, in our experiment, we would need more judges to be able to generalize the analysis and then we would need to enlarge our study by increasing the number of judges, and new data.

In our opinion, the question "Is my judge a good one?" can be answered nevertheless. There is no such thing as a perfect judge, someone who is better thean the others. We should rather rely on several judges who will be enough to build a reliable evaluation with the combination of their judgments.

6. References

Hervé Blanchon, Christian Boitet, Francis Brunet-Manquat, Mutsuko Tomokiyo, Agnès Hamon, Vo Trung Hung, and Youcef Bey. 2004. Towards Fairer Evaluations of Commercial MT Systems on Basic Travel Expressions Corpora. In *Proceedings of IWSLT 2004 (ICLSP 2004 Satellite Workshop)*, pages 21–26, Kyoto, Japan, September–October.

Chris Callison-Burch, Cameron Shaw Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (Meta-) Evaluation of Machine Translation. In *Proceedings of Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic, June.

John B. Carroll. 1966. An experiment in evaluating the quality of translations. *Mechanical Translation and Computational Linguistics*, 9(3 and 4):55–66, September–December.

Matthias Eck and Chirori Hori. 2005. Overview of the IWSLT 2005 Evaluation Campaign. In *International Workshop on Spoken Language Translation: Evaluation Campaign on Spoken Language Translation*, Pittsburgh, PA, USA, October.

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

Olivier Hamon, Anthony Hartley, Andrei Popescu-Belis, and Khalid Choukri. 2007. Assessing Human and Automated Quality Judgments in the French MT Evaluation Campaign CESTA. In *Proceedings of the MT Summit XI*, Copenhagen, Denmark, September.

Olivier Hamon, Djamel Mostefa, and Victoria Arranz. 2008. Diagnosing human judgments in MT evaluation: an example based on the Spanish language. In *Proceedings of MATMT 2008: Mixing Approaches to Machine Translation*, pages 19–26, San Sebastian, Spain, February.

Philipp Koehn. 2007. Evaluating Evaluation Lessons from the WMT 2007 Shared Task. In *Proceedings of the MT Summit XI Workshop on Automatic Procedures in Machine Translation Evaluation*, Copenhagen, Denmark, September.

J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174, March.

Keith J. Miller and Michelle Vanni. 2005. Inter-rater agreement measures, and the refinement of metrics in the PLATO MT evaluation paradigm. In *Proceedings of the MT Summit X*, pages 125–132, Phuket, Thailand, September.

Georges van Slype. 1979. Critical Study of Methods for Evaluating the Quality of Machine Translation. Technical Report Final report BR 19142, Brussels: Bureau Marcel van Dijk.

John S. White, Theresa A. O'Connel, and Francis E. O'Mara. 1994. The ARPA MT evaluation methodologies: evolution, lessons, and future approaches. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas*, pages 193–205, Columbia, Maryland, USA, October.

Yang Ye and Steven Abney. 2006. How and Where do People Fail with Time: Temporal Reference Mapping Annotation by Chinese and English Bilinguals. In *Workshop on Frontiers in Linguistically Annotated Corpora 2006*, pages 13–20, Sydney, Australia, July.