# The FAUST Corpus of Adequacy Assessments for Real-World Machine Translation Output

**Daniele Pighin, Lluís Màrquez, Lluís Formiga**

Universitat Politècnica de Catalunya - Barcelona, Spain
{pighin,lluism}@lsi.upc.edu, lluis.formiga@tsc.upc.edu

## Abstract

We present a corpus consisting of 11,292 real-world English to Spanish automatic translations annotated with relative (ranking) and absolute (adequate/non-adequate) quality assessments. The translation requests, collected through the popular translation portal http://reverso.net, provide a most variated sample of real-world machine translation (MT) usage, from complete sentences to units of one or two words, from well-formed to hardly intelligible texts, from technical documents to colloquial and slang snippets. In this paper, we present 1) a preliminary annotation experiment that we carried out to select the most appropriate quality criterion to be used for these data, 2) a graph-based methodology inspired by Interactive Genetic Algorithms to reduce the annotation effort, and 3) the outcomes of the full-scale annotation experiment, which result in a valuable and original resource for the analysis and characterization of MT-output quality.

**Keywords:** Machine Translation; Quality Assessments; Annotated Corpus

## 1. Introduction

Computational approaches to confidence and quality estimation (QE) for machine translation (MT) have been receiving increasing attention in the last decade, e.g., Blatz et al. (2004), Specia et al. (2009), Pighin and Màrquez (2011), Banchs and Li (2011), and the reasons for this fact are very practical: being able to measure the performance of translation models, without having to manually produce hundreds or thousands of reference translations, is a necessary step in order to compare the output of alternative systems and to deploy adequate and reliable automatic translation products to a broader audience.

Models of translation quality are generally learned using manual annotations collected for MT evaluation campaigns, such as the yearly editions of the Workshop on Machine Translation (WMT), e.g., Koehn and Monz (2006), Callison-Burch et al. (2010), or the collection of postediting-effort assessment compiled by Specia et al. (2010). While all these datasets are extremely valuable resources, being mostly based on transcriptions of European Parliament sessions (Koehn, 2005) (and, to a lesser extend, news-wire data) the quality models that we can learn from them are hardly adaptable to domains in which language is less structured, possibly noisy and less predictable.

In this paper, we present a dataset of 11,292 English-to-Spanish, real-world translations annotated with relative and absolute quality assessments, collected in the context of the *Feedback Analysis for User adaptive Statistical Translation* (FAUST) EU project[1]. FAUST focuses on the development of machine translation systems that can respond rapidly and intelligently to user feedback, and as such it is centered around user provided translation requests and feedback. Unlike already available resources, the data in this corpus reflects the real needs and requirements of casual users of translation systems, and covers a wide spectrum of domains and styles. Some requests are complete, well-formed sentences, whereas others are just snippets of text copied and pasted from somewhere else (e.g., chat rooms, web pages, software manuals, just to name a few), or simply words or noun phrases in isolation. In many cases, the input is disfluent or ungrammatical. In some cases, the interpretability of the input sentence is questionable. Nevertheless, real-world translation systems must be able to cope with this kind of data, and to produce outputs which, at the very least, should contain useful clues to satisfy practical needs of users.

In order to deliver a high-quality resource, we implemented an annotation strategy that attempts to reduce the fatigue, ambiguity and frustration involved in long annotations tasks, which lead to contradictions (noise) in the annotation process (Takagi, 2001). These aspects have been formerly identified on the field of Interactive Evolutionary Computation, where high-repetitive evaluation tasks are extremely common. To limit annotation noise, we apply a hierarchical, graph-based annotation model to user annotations (Llorà et al., 2005). This graphical approach makes it possible to discard noisy users or sentences, split the annotation task into a collaborative multi-user task, identify the ambiguity of specific source/target pairs or obtain a complete ranking of different translation systems from simple pairwise comparisons.

The rest of the paper is structured as follows: in Section 2 we outline the steps of the annotation process; in Section 3 we describe a preliminary experiment that we conducted to select the most appropriate quality criterion for the FAUST dataset; in Section 4 we describe the annotation of the dataset in terms of relative assessments (rankings), whereas in Section 5 we describe its annotation in terms of absolute assessments (adequate/non adequate); finally, in Section 6 we draw our conclusions.

[1] FP7-ICT-2009-4, http://www.faust-fp7.eu/faust

## 2. Annotation overview

Our objective is to build a corpus of rankings *and* absolute quality annotations for alternative translations of the same source sentences. These two layers of annotation are complementary and useful in different ways, and they can be exploited to learn models of quality with different applications, i.e., to select among alternative translations or to discard unsatisfactory outputs.

We considered 1,882 translation requests in English submitted to *Softissimo*'s online translation portal, `http://www.reverso.net`. Softissimo is one of the technological partners of the FAUST project. A professional translator corrected the most obvious typos and provided reference translations into Spanish for all of them. The corrected sentences have been automatically translated into Spanish with five different systems: two of them are provided by partners of the FAUST project (one by Language Weaver, the other by the TALP team at UPC); the remaining three systems are on-line commercial systems that we queried via their web APIs, namely *Google Translate*[2], *Bing Translator*[3] and *Systran*[4].

The annotation activity has been organized around three tasks:

1. A preliminary experiment aimed at selecting the most appropriate quality criterion for the dataset;
2. The annotation of pairwise comparisons by using a graphical model that allows us to obtain a fully ranked set of translations with minimum effort;
3. The annotation of each translation with an absolute quality assessment (adequate/non adequate) also based on the graphical model.

These steps are fully documented in the following sections.

## 3. Selection of an annotation criterion

We conducted a preliminary annotation acitivity with the twofold objective of 1) comparing different criteria for pairwise ranking of translation hypotheses, and 2) understanding what aspects of the annotation task can more easily result in annotation bias. In other words, we were trying to understand 1) which criterion can more easily be related to a person's idea of translation quality, and 2) which one is more likely to produce more stable and objective results.

### 3.1. Methodology

The annotators were asked to annotate 89 triplets of sentences according to five different translation quality criteria. Each triplet is in the form $\langle s, h_1, h_2 \rangle$, where $s$ is a source sentence (English) and $h_1$ and $h_2$ are two translation hypotheses for $s$ (Spanish). For this task we selected a combination of sentences from the FAUST data and Europarl (Koehn, 2005) data, with the idea of understanding how different translation criteria relate to sentences of different length, nature and complexity. As for the translation hypotheses, we also included in the data some reference translations to be used as a quality check.

| Criterion | Avg($t'$) | Dev($t'$) |
|---|---|---|
| Fluency | 0.18 | 0.24 |
| Post-edit effort | 0.20 | 0.24 |
| Adequacy | 0.20 | 0.24 |
| Adequacy+Fluency | 0.21 | 0.26 |
| Goodness | 0.22 | 0.23 |

Table 1: Aggregate normalized annotation time (average and standard deviation) for the five annotation criteria.

For each triplet, all the annotators were required to select which of the two hypotheses, according to a specific quality criterion, is a better translation of the source sentence. The annotators were instructed to select one of the two hypotheses only if the difference between them was noticeable. Otherwise, they were invited to mark the two hypotheses as equivalent.

The five quality criteria were defined as follows:

- **Fluency** - Translation fluency (i.e. ”which translation reads better, is more grammatical and cohesive?”)
- **Adequacy** - Translation adequacy (i.e. ”which translation conveys more exactly the information of the source sentence?”)
- **Adequacy+Fluency** - A combination of adequacy and fluency (i.e. ”which translation is better, both in terms of the amount of information it exactly conveys and its grammaticality/correctness?”)
- **Post-edit effort** - An estimate of required post-editing effort (i.e. ”which translation would be easier to edit in order to get a publication-ready sentence?”)
- **Goodness** - A subjective measure of translation quality (i.e. ”which translation seems better?”).

Some annotators have pointed out that it was very difficult to differentiate between “Goodness” and “Adequacy+Fluency”. The annotation guidelines clearly emphasized that the former should have been a completely subjective measure, more related to personal taste and not necessarily easy to quantify with respect to adequacy and/or fluency. For example, the annotator may have wished to select translations that sound more human-generated than machine-generated, or freer vs. more literal translations.

#### 3.1.1. Difficulty of the annotation process

The annotation of each triplet was timed, as annotation time can be regarded as an objective indicator of the difficulty of the task. In an attempt to reduce the bias introduced by the order in which the criteria were applied (i.e., users tend to be faster after observing the same triplets several times) in the reminder all annotation times are normalized according to the formula:

$$t' = \frac{t - \min(t)}{\max(t) - \min(t)} \, ,$$

[2] `http://translate.google.com`
[3] `http://www.microsofttranslator.com`
[4] `http://systransoft.com`

| UID | Criterion | Avg | Dev |
|---|---|---|---|
| 2 | Adequacy+Fluency | 0.15 | 0.23 |
| | Fluency | 0.15 | 0.24 |
| | Adequacy | 0.16 | 0.23 |
| | Goodness | 0.18 | 0.18 |
| | Post-edit effort | 0.21 | 0.22 |
| 3 | Adequacy | 0.17 | 0.22 |
| | Fluency | 0.18 | 0.25 |
| | Adequacy+Fluency | 0.20 | 0.26 |
| | Post-edit effort | 0.24 | 0.28 |
| | Goodness | 0.25 | 0.25 |
| 4 | Goodness | 0.14 | 0.20 |
| | Post-edit effort | 0.15 | 0.19 |
| | Adequacy | 0.20 | 0.20 |
| | Fluency | 0.22 | 0.26 |
| | Adequacy+Fluency | 0.30 | 0.28 |
| 5 | Adequacy | 0.15 | 0.21 |
| | Adequacy+Fluency | 0.17 | 0.25 |
| | Fluency | 0.20 | 0.26 |
| | Post-edit effort | 0.22 | 0.27 |
| | Goodness | 0.25 | 0.26 |
| 6 | Fluency | 0.15 | 0.17 |
| | Post-edit effort | 0.19 | 0.20 |
| | Adequacy+Fluency | 0.23 | 0.25 |
| | Goodness | 0.27 | 0.24 |
| | Adequacy | 0.33 | 0.30 |

Table 2: Per-user normalized annotation time (average and standard deviation) for the five annotation criteria.

where $t'$ is the normalized annotation time, $t$ is the observed time and $\max(t)$ and $\min(t)$ are the maximum and minimum annotation time for each user when using each criterion, respectively. $t'$ is an indicator of how annotations tend to be generally faster ($0 \leq t' \ll 1$) or slower ($0 \ll t' \leq 1$). Per-triplet annotation times larger than 300 seconds (5 minutes) have been ascribed to temporary interruptions in the annotation activity and therefore are not included in the stats.

Table 1 shows average and standard deviation for the normalized annotation time for each criterion. As expected, annotating the fluency of translation is a simpler task. In fact, of the five criteria fluency is the simplest to understand and to apply, being only based on surface features of the sentence. Post-edit effort and adequacy require very similar effort, whereas trying to account for more aspects of the translation at the same time (as in the case of adequacy+fluency and goodness) results in slightly more difficult annotations.

Table 2 breaks down annotation time for each annotator with respect to each of the five criteria. Even though it is difficult to establish a clear pattern across all five annotators, we can still observe that, with the exception of Annotator 4, all the other annotators have ranked fluency among the three easiest criteria, and goodness among the 2 most difficult.

(a) All data (Europarl + FAUST, 89 triplets)

| Criterion | MC-5 | MC-4 | MC-3 | UND | Ties |
|---|---|---|---|---|---|
| Fluency | 23.60 | 28.09 | 42.70 | 5.62 | 6.76 |
| Adequacy | 37.08 | 32.58 | 26.97 | 3.37 | 5.17 |
| Adequacy+Fluency | 30.34 | 24.72 | 37.08 | 7.87 | 8.86 |
| Goodness | 28.09 | 28.09 | 37.08 | 6.74 | 8.11 |
| Post-edit effort | 23.60 | 23.60 | 43.82 | 8.99 | 9.76 |

(b) Europarl data (35 triplets)

| Criterion | MC-5 | MC-4 | MC-3 | UND | Ties |
|---|---|---|---|---|---|
| Fluency | 20.00 | 28.57 | 48.57 | 2.86 | 3.45 |
| Adequacy | 37.14 | 34.29 | 22.86 | 5.71 | 8.00 |
| Adequacy+Fluency | 34.29 | 22.86 | 42.86 | 0.00 | 0.00 |
| Goodness | 22.86 | 40.00 | 31.43 | 5.71 | 6.25 |
| Post-edit effort | 31.43 | 17.14 | 45.71 | 5.71 | 6.06 |

(c) FAUST data (54 triplets)

| Criterion | MC-5 | MC-4 | MC-3 | UND | Ties |
|---|---|---|---|---|---|
| Fluency | 25.93 | 27.78 | 38.89 | 7.41 | 8.89 |
| Adequacy | 37.04 | 31.48 | 29.63 | 1.85 | 3.03 |
| Adequacy+Fluency | 27.78 | 25.93 | 33.33 | 12.96 | 15.22 |
| Goodness | 31.48 | 20.37 | 40.74 | 7.41 | 9.52 |
| Post-edit effort | 18.52 | 27.78 | 42.59 | 11.11 | 12.24 |

Table 3: Inter-annotator agreement for each of the five criteria. The columns report the percentage of annotations in each majority class.

### 3.1.2. Inter-annotator agreement

The annotation was carried out by five annotators. Four of them are native Spanish speakers, with adequate command of English. One of these has very good command of English. The fifth annotator is a native American English speaker, with very good command of Spanish.

Table 3 shows the inter annotator agreement in terms of majority classes. The results are provided for the whole dataset, in Table 2(a), for Europarl data only, in Table 2(b), and for FAUST data only, in Table 2(c). A majority class of 5 (MC-5) means that all the five annotators have annotated a sentence in the same way. Similarly, a majority class of 3 (MC-3) marks the cases in which 3 annotators out of 5 have taken the same decision. The column labeled ”UND” shows the percentage of cases in which no decision can be taken (i.e. there are two options with two preferences and one option with just one preference). The column labeled ”Ties” show the percentage of cases in which the majority of annotators indicated that two hypotheses are equivalent. The results on the whole dataset, presented in Table 2(a), show that adequacy-based annotations tend to be more consistent: using this criterion, in 37% of the cases all the annotators agree on their decision, and in 32% of the cases only one annotator disagrees; only in 3% of the cases the annotators cannot reach an agreement (UND). This figures are noticeably better than for the all the other criteria. Especially concerning post-editing effort, the agreement among annotators is relatively low. As one of the annotators pointed out, the application of this criterion is complicated by the fact that the two alternatives can be edited in many ways, and to decide about the relative quality the annotator has to imagine one specific editing scenario for each sentence. These choices then have an effect on the final decision.

Adequacy enforces higher agreement also when we consider the results obtained on Europarl or FAUST data separately, as shown in tables 2(b) and 2(c), respectively. By comparing them, we can observe that on FAUST data, consisting for the most part of very short or ungrammatical sentences, adequacy also results in a very small number of undecided cases (1.85%), and only in very few cases (3.03%) the annotators cannot rank the two hypotheses. This might be related to the fact that the lack of a well-formed sentence structure makes it easier for the annotators to focus on aspects related to semantics without being distracted by the fluency of the output.

The evidence collected motivated us to select adequacy as a translation quality criterion for the full-scale annotation task. In fact, adequacy-oriented annotation is only slightly more difficult (annotation-time wise) than fluency, even though not as difficult as Post-edit effort. On the other hand, it also results in the most consistent annotations, and is especially reliable when employed to rank translations of short or non-fluent inputs, which are very frequent in the FAUST dataset.

## 4. Relative adequacy assessments

For the full-scale annotation experiment, we reused the same setting as before (pairwise translation ranking), but focused on the alternative translations for the 1,882 FAUST source sentences using adequacy as the quality criterion. In the annotation guidelines, the annotators were invited to opt for the "tie" decision in all cases in which, according to their judgement, they would be equally satisfied (or dissatisfied) upon being offered any of the two translations.

### 4.1. Methodology

In order to overcome the problems of the perceptual noisy annotation and minimize the number of triplets to be annotated, we used applied a hierarchical graph based scheme inspired by active Interactive Genetic Algorithms (aiGAs) (Llorà et al., 2005). In the context of natural language processing, a similar methodology has already been adopted by Formiga et al. (2010) for the task of weight tuning in unit selection for speech synthesis.

The method requires the annotators to annotate just enough triplets to build a connected graph $\mathcal{G} = \langle V, E \rangle$ of all translation alternatives $V$ for the same source sentence, as shown in Figure 1 (1). Based on the decisions of the annotators, we can build a second graph (2) whose edges $E$ represent pairwise comparisons between translation alternatives $V$. The edges can be either directed (i.e., the user has preferred one translation over the other) or undirected (i.e., the two translations are equivalent). This connected, partially undirected graph can easily be turned into a fully-directed graph $\mathcal{G}'$, by exploiting topological properties of the graph deriving from its construction. The process also eliminates loops and inconsistencies from the graph (3) and collapses into a single node all the equivalent translations (4). Finally, the global ranking of all the vertices $V$ can be obtained by sorting them according to the calculated *dominance* of each vertex $\hat{f}(v) = \delta(v) - \phi(v)$, where $\delta(v)$ is the number of vertices that $v$ dominates and $\phi(v)$ is the number of vertices that dominate $v$. In Figure 1 (4) the dominance of each node is shown next to it. For this case, the final ranking would be:

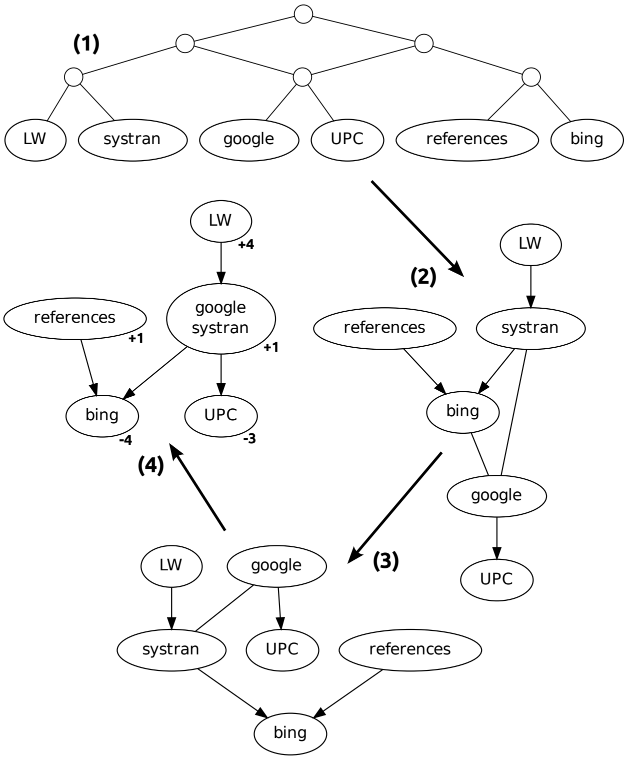

Figure 1: Tournament-based translation ranking.

1. LW, with $v = 4$;
2. Google, Systran, References, with $v = 1$;
3. UPC, with $v = -3$;
4. Bing, with $v = -4$.

By following this scheme, the number of pairwise comparisons necessary to establish a complete rank of the alternative translations for a sentence is reduced from $\sim 6 \log_2 6 = 15.5$ to 6, i.e. the number of required annotations is decreased by more than 60%. To further streamline the process, we asssume any two translations which differ only in the casing of words to be equivalent. As a result, we are left with 10,203 triplets requiring explicit annotation instead of ∼29,155, i.e. total annotation effort is reduced by approximately 65%.

### 4.2. Inter-annotator agreement

This annotation was carried out by 16 annotators, all of them researchers at UPC's TALP center and native Spanish speakers. We set apart 10 source sentences (60 triplets) to measure the inter-annotator agreement. These sentences have been selected so as to build a varied and representative selection of FAUST data:

1. A technical heading: *Modeling Roundabout Traffic Flow as a Dynamic Fluid System*;

| MC | Frequency (%) | | |
|---|---:|---|---|
| 7 | 6.78 | relative majority (20.34%) | |
| 8 | 13.56 | | |
| 9 | 13.56 | absolute majority (79.66%) | |
| 10 | 11.86 | | |
| 11 | 11.86 | | more than 2/3 (54.24%) |
| 12 | 10.17 | | |
| 13 | 6.78 | | |
| 14 | 13.56 | | |
| 15 | 11.86 | | |

Table 4: Majority classes on the pairwise annotation task and their frequencies.

2. A very short, polysemous passage: *too smart*;

3. A domain-specific (commercial) text: *Thanks for your purchase of ScanSoft voices from NextUp.com. These voices are intended for use within TextAloud and other NextUp.com products*;

4. A conversational snippet: *I don't want to say that you are wrong with your religion, you can believe it... no problem. I just want to know, what do you think about the other religions*;

5. An MT-generated sentence (from German): *Welcome to our new gym in Konitsa! A change in your life plays in many people a large role. It would develop new, dynamic start.... der summer!!! One of the best seasons for achieving objectives. The summer strengthens us with his lebensfreude…*;

6. An ungrammatical segment, either the product of SMT or written by a non-native speaker with poor English command: *My good morning love, Everything well? Yesterday I fell asleep excuse. I find a lot albeit you have wakened up to order me message, I am to play*;

7. A passage with many non-linguistic tokens (like a copied and pasted web-form): *Vendor Full Name: Channel (Local/WIU/WIJ/WTO): WIU (Name & Sign) Requested by : Date :2010.1.26.*;

8. A rather unstructured list of localities: *Great Wall of China, China[7] The Taj Mahal, India[8][9] Stonehenge, United Kingdom[10] Machu Picchu, Peru[11] Banaue Rice Terraces, Philippines[12][13][14][15][16] The Terracotta Army of Xi'an, China.[17][18] Amber Room in the Catherine Palace near Saint Petersburg, Russia[19] The monastery of San Lorenzo del Escorial, Spain.[20]*;

9. A technical biology text: *the addition of cholera toxin represented heavily confluent multilayered cultures. Normal cells maintained in cholera toxin showed rapid growth in 3rd passage, although with less piling up of cells at 8 to 10 d. Cells mainrained without cholera toxin showed poor growth in 3rd passage and did not achieve confluence*;

10. A crude and vulgar sentence, with slang and swear words *[omitted due to its possibly offensive content].*

These sentences have been annotated by all the 16 annotators. All the other sentences have been annotated by only one annotator. To increase consistency within the annotation, all the triplets relative to the same source sentence are assigned to the same annotator.

Table 4 shows the relative frequency of the majority classes (MC) observed on the pairwise ranking annotation task. Only in 20.34% of the cases the most popular option was not selected by the absolute majority of annotators (i.e., 9). The absolute majority of the annotators agreed on the decision in almost 80% of the cases, and in almost 55% of the cases at least 2 out of 3 annotators took the same decision. Cohen's $\kappa$, measured between the two most prolific annotators, is $\kappa$=0.55. These figures show that, even though the task is a difficult one, the problem definition is precise enough to allow for a good inter-annotator agreement. Especially in the light of the heterogeneity of the inter-annotator set, the figures confirm the accuracy of the annotation process.

Concerning decision classes, the "tie" option is the most popular, being selected in approximately 47% of the cases. This behaviour is in line with our expectations, as we are interested in telling two translations apart only in those cases in which it can have an effect on the perceived quality of translation, and the annotators were instructed accordingly.

## 5. Absolute adequacy assessments

This last annotation activity was aimed at annotating each of the 11,292 translations in the FAUST corpus with binary, absolute quality assessments (adequate/non adequate).

### 5.1. Methodology

To reduce annotation effort, we assume that the reference translation and all the translations dominating it are automatically adequate. In the example of Figure 1, the reference translation is not dominated by any other node, and therefore only the reference translation would be automatically considered adequate. As shown in the first block of rows of Table 5, just building the graphs and relying on reference dominance allowed us to cut required annotations from 11,292 to 4,298. Subsequently, for each graph we rank the remaining nodes based on the number of dominated nodes $\delta(v)$ (see Section 4.1) and ask the annotators to

| | |
|---|---|
| Absolute assessments in the dataset | 11,292 |
| Non-trivial annotations (after graph building) | 4,298 |
| Effort reduction (%) | 61.94 |
| Actual annotations done | 3,862 |
| **Actual effort reduction (%)** | **65.80** |

Table 5: Absolute assessments effort reduction thanks to the graphical approach.

mark each translation as adequate or not. Whenever a translation is marked as non-adequate, all the dominated translations can automatically be marked as inadequate. Conversely, when a translation is marked as adequate we cannot assume anything about the quality of lower-ranked translations, and the dominated nodes have to be annotated separately. By exploiting dominance we further reduced the annotation effort from 4,298 annotations to 3,862. With respect to the goal of annotating 11,292, exploiting the topological properties of the graphs allowed us cut the effort to annotate the translations with absolute adequacy assessments by almost 66%.

### 5.2. Inter-annotator agreement

This activity was carried out by two native Spanish speakers with good command of English. Inter-annotator agreement was calculated on a shared set of 66 translations obtained from 15 randomly selected source sentences. Cohen's $\kappa$ between the two annotators is 0.56, Pearson's correlation is 0.61 and Spearman's correlation is 0.45. All these indicators show a substantial agreement between the annotators.

### 5.3. Harmonization and cross-method agreement

The directed graphs that we use as the basis for the absolute adequacy annotation are not completely connected, as shown by the example in Figure 1 (4). As a consequence, it may happen that a higher-ranked translation is marked as inadequate, whereas a lower ranked translation is marked as adequate. For example, with respect to Figure 1 (4), the annotator may label the translation provided by UPC (ranked 3rd) as inadequate, and the translation by Bing (ranked 4th) as adequate.

To overcome such apparent inconsistencies, in the annotated dataset we provide two versions of the ranks:

- The original, dominance-based ranks;
- A harmonized version of the ranks, in which the ranks have been re-sorted so that all the adequate translations are never ranked lower than an inadequate translation.

This post-processing step also allows us to estimate the agreement between the two annotation stages, which we can measure by means of the correlation between the original and the harmonized ranks. To this end, we measured the Spearman correlation, the Mean Absolute Error and the Root Mean Squared Error between the two ranks for each sentence, and then averaged these values. The results, listed in Table 6, show very high Spearman correlation and negligible differences in the ranks, and demonstrate the consistency among the different stages of the annotation process.

| Method | Avg | Dev |
|---|---|---|
| Spearman correlation | 0.98 | 0.06 |
| Mean Absolute Error (MAE) | 0.05 | 0.17 |
| Root Mean Squared Error (RMSE) | 0.08 | 0.23 |

Table 6: Agreement between pre- and post-reordering ranks.

## 6. Conclusions

We presented a dataset consisting of 11,292 relative and as many absolute adequacy assessments for real-world machine-translation output. We detailed a preliminary experiment in which we selected adequacy as the most appropriate quality criterion for the task, and we outlined an effective methodology to reduce the fatigue in large-scale annotation exercises.

The full body of annotations is available for download at `ftp://mi.eng.cam.ac.uk/data/faust/-UPC-Oct11-FAUST-quality-assessments.tgz`. We share it with the research community under a Creative Commons license, in the hope that it will foster research in quality estimation for machine translation and constitute a valuable asset for comparative studies of the quality indicators characterizing different typologies of data.

## Acknowledgements

We would like to express our gratitude to all the friends and colleagues at TALP who made this work possible by participating in the annotation process. This research has been partially funded by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement numbers 247762 (FAUST project, FP7-ICT-2009-4-247762) and 247914 (MOLTO project, FP7-ICT-2009-4-247914) and by the Spanish Ministry of Education and Science (OpenMT-2, TIN2009-14675-C03).

## 7. References

Rafael E. Banchs and Haizhou Li. 2011. AM-FM: A Semantic Framework for Translation Quality Assessment. In *Proceedings of ACL'11*, pages 153–158, Portland, Oregon, USA, June. Association for Computational Linguistics.

John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2004. Confidence estimation for machine translation. In *Proceedings of the COLING'04*, Stroudsburg, PA, USA. ACL.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, and Omar Zaidan, editors. 2010. *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*. ACL, Uppsala, Sweden.

L. Formiga, F. Alías, and X. Llorà. 2010. Evolutionary Process Indicators for Active IGAs Applied to Weight

Tuning in Unit Selection TTS Synthesis. In *Proceedings of the IEEE CEC'10*, Barcelona.

Philipp Koehn and Christof Monz, editors. 2006. *Proceedings of the Workshop on Statistical Machine Translation*. Association for Computational Linguistics, New York City.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

Xavier Llorà, Kumara Sastry, David E. Goldberg, Abhimanyu Gupta, and Lalitha Lakshmi. 2005. Combating User Fatigue in iGAs: Partial Ordering, Support Vector Machines, and Synthetic Fitness. In *Proceedings of GECCO'05*, pages 1363–1370, New York, NY, USA.

Daniele Pighin and Lluís Màrquez. 2011. Automatic projection of semantic structures: an application to pairwise translation ranking. In *Fifth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-5)*, Portland, Oregon. ACL.

Lucia Specia, Marco Turchi, Zhuoran Wang, John Shawe-Taylor, and Craig Saunders. 2009. Improving the Confidence of Machine Translation Quality Estimates. In *Proceedings of Machine Translation Summit XII*, Ottawa, Canada.

Lucia Specia, Nicola Cancedda, and Marc Dymetman. 2010. A Dataset for Assessing Machine Translation Evaluation Metrics. In *Proceedings of LREC'10*, Valletta, Malta. European Language Resources Association (ELRA).

H. Takagi. 2001. Interactive Evolutionary Computation: Fusion of the Capabilities of the EC Optimization and Human Evaluation. *Proceedings of the IEEE*, 89(9):1275–1296.