All Theses and Dissertations

2012-05-25

Ruqual: A System for Assessing Post-Editing

Jason K. Housley

Brigham Young University - Provo

Follow this and additional works at: https://scholarsarchive.byu.edu/etd

Part of the Linguistics Commons

Ruqual: A System for Assessing Post-Editing

Jason Keith Housley

A selected project submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements of the degree of

Master of Arts

Alan K. Melby, Chair
Deryle W. Lonsdale
Mark E. Davies

Department of Linguistics and English Language

Brigham Young University

June 2012

ABSTRACT

Ruqual: A System for Assessing Post-Editing

Jason Keith Housley
Department of Linguistics and English Language
Master of Arts

Post-editing machine translation has become more common in recent years due to the increase in materials requiring translation and the effectiveness of machine translation systems. This project presents a system for formalizing structured translation specifications that facilitates the assessment of the performance of a post-editor. This report provides details concerning two software applications: the Ruqual Specifications Writer, which aids in the authoring of post-editing project specifications, and the Ruqual Rubric Viewer which provides a graphical user interface for filling out a machine readable rubric file. The project as a whole relies on a definition of translation quality based on the specification approach.

In order to test whether potential evaluators are able to reliably assess the quality of post-edited translations, a user study was conducted that utilized the Specifications Writer and Rubric Viewer. The specifications developed for the project were based on actual post-editing data provided by Ray Flournoy of Adobe. The study involved simulating the work of five post-editors, which consisted of developing texts and scenarios. 17 non-expert graders rated the work of the five fictional post-editors, and an Intraclass Correlation of the graders responses shows that they are reliable to a high degree. The groundwork laid by this project should help in the development of other applications that assist in the assessment of translation projects in terms of a specification approach to quality.

ACKNOWLEDGEMENTS

Table of Contents

List of Tables

List of Figures

1. INTRODUCTION. The progress of globalization has produced an ever increasing quantity of materials requiring translation. Moreover, there is an insufficient number of highly skilled translators to handle the burden of translatable materials. In other words, the world has moved beyond the assessment of the ALPAC report, which argued that there were more than enough translators to meet the demand of government translation (Hutchins 2007), into an era where the need for translation has made services like Google Translate widespread. Furthermore, advances in the processing power of computers and the recent successes of statistical approaches to machine translation (MT) have increased opportunities for developing practical applications of MT. One practical application of MT that helps to allow it to be applied to a variety of situations is the addition of a human editor who corrects the raw MT output to meet a particular set of requirements that the MT system would not be able to meet on its own. This process is called POST-EDITING and has become increasingly more common in recent years. However, post-editing presents its own set of problems and challenges, not the least of which is the question of how to determine whether a person has done an adequate job as a post-editor.

Post-editing provides an alternative to either accepting often less than adequate raw MT output or paying the often prohibitive cost of a human translator. However, post-editing ultimately requires the labor of a human editor who must be selected or trained to meet the requirements of the post-editing project. If finding or training a post-editor proves to be a significant burden, then it may be more cost effective for companies to hire the services of a professional translator who can render the translation without relying on raw MT output. Post-editing cuts costs most effectively when the MT output is already generally acceptable and the post-editor makes as few changes as possible or spends the least amount of time necessary to produce the final product. This means that highly skilled translators, who understand the source

text and have a clear idea of how the language ought to be rendered in the target text, may actually prove to be quite unacceptable post-editors because they will often correct too much of the raw MT output. Spending too much time on a particular text reduces the cost-effectiveness of the post-editing. On the other hand, even in cases where the target text does not need to be highly accurate in respect to the source text, a certain degree of proficiency in the source language is often required on the part of the post-editor. A post-editor is therefore more than simply someone who corrects the grammar of the target text, but at the same time they are not required to fully translate the text. This balance may be difficult to obtain.

Furthermore, the actual requirements of the post-editing project often vary from project to project compounding the difficulty of assessing the work of the post-editor. For one project, a post-editor may be asked simply to modify the sentence structure of the raw MT output so that it is easier to read and understand, but the target text may still contain obvious indications that it is a translation. In another project, the post-editor may be asked to modify the MT output in such a way that the target text appears to be originally written in the target language even if this means removing or changing more elements in the raw MT output. In order to evaluate projects with disparate requirements, we either need to have a series of specialized methods or an expert evaluator trained in assessing post-editing of the subject matter in question, which would lead to a greater burden of management for translation companies; or we need a method for assessing the quality of post-editing projects that is flexible enough to handle these differences, consistent enough to provide a reliable result, and simple enough that the burden of training an evaluator is worthwhile.

In addition, the task of evaluating the work of individuals naturally raises questions about the reliability of those conducting the evaluation. A major focus of this project is the question of

reliability among potential evaluators of post-editing. Specifically, the research question for the project is as follows:

> How reliably can non-expert human graders assess the quality of post-edited machine translation, when the initial English target text was generated by a free and publically available machine translation system, when the source text is a medium difficulty (ILR level 2) document in Japanese, and when the graders assessing the performance of the post-editors are given a rubric based on a set of structured translation specifications?

This research question may not be immediately intuitive at this point, but the particulars will appear more relevant as the rest of this project is explained. This research question was explored via a user study simulating the assessment of post-editors by real human graders using software tools provided in this project.

This project addresses some of the challenges associated with post-editing by providing a software application that generates a rubric for assessing a post-editor. Since this project uses a rubric to assess quality, the name of the project is Ruqual, which is a blend of "rubric" and "quality." The rubric and overall system relies on the framework for developing structured specifications for translation projects provided in ISO/TS 11669 (General Guidance -- Translation Projects). Using structured specifications provides a consistent framework for evaluating almost all translation projects while allowing for the flexibility necessary to handle the broad variation in the circumstances in which translation may be carried out. Moreover, the use of a rubric should allow for non-expert evaluators to reliably assess the work of a given post-editor. This report provides the details of the Ruqual project and a user study verifying that, at least in the case tested, non-expert evaluators are able to reliably assess a number of post-edited translations.

Chapter 2 of this report will detail relevant background research relating to defining translation quality, previous work in assessing post-editing, and rubrics as a tool for translation assessment. Chapter 3 will explain the project design and software implementation. In Chapter 4 the user study design will be presented, and Chapter 5 will provide an analysis of the data obtained from the user study. Chapter 6 will provide a discussion and conclusion of the project findings, and Chapter 7 will be dedicated to exploring options for future work.

2. BACKGROUND. The idea of post-editing machine translation (MT) dates back to the inception of MT where it was thought that a pre-editor and a post-editor may be required to adjust the textual input and output of the machine respectively (Booth & Booth 1953). However, this practice has only recently become practical as a commercial activity with the rise in materials requiring less-than-publishable quality and more capable machine translation hardware and software. Translated materials of less-than-publishable quality may be required for various purposes such as a customer service knowledge base for internal use in troubleshooting various software programs. Evaluating post-editing presents challenges similar to evaluating more general translation. This project proposes to use the SPECIFICATION APPROACH (Hague et al. 2011) to assessing quality, in accordance with the recent trend of creating standards in the area of translation (Melby 2011). However, some differences between post-editing and general translation have led to some post-editing specific methods for evaluation. This is apparently the first rubric-based approach intended strictly for evaluating post-editing although prior work has established rubrics designed for general translation and various error rate measures used in translation proficiency testing (Hague et al. 2011). Section 1 of this chapter will discuss the definition of translation quality espoused in this project and relate it to some established standards of translation quality. Section 2 will focus on past work on the evaluation of post-editing and other rubric-based approaches to assessing a translation.

2.1 DEFINITION OF QUALITY. This project's definition of translation quality goes beyond the common definition of a good translation: one that is both accurate and fluent. Accuracy generally means that the target text corresponds precisely to the source text, whereas fluency describes how natural, acceptable, or grammatically correct the target text is in the target language. I will call this perspective the TRANSCENDENT APPROACH to quality; in this perspective, the more

natural and accurate a target text is, the higher its quality. However, this definition of quality is impractical for real translation projects because it makes the implicit assumption that translation quality is absolute. The transcendent approach assumes that a given target text having a high degree of accuracy and fluency will be a good translation in all circumstances. This definition also implies that accuracy and fluency can be achieved at a high degree in all cases, but in fact they often undercut one another.

In order to obtain a high degree of accuracy, the translator often has to sacrifice some fluency. This is especially true when translating between language pairs that are very different. Likewise, the translator may sacrifice some accuracy in order to make a translation sound better in the target language. If one were to accept an absolute, transcendent view of quality, there would be no way for the translator to reliably determine which of the two sides is more important other than via his or her own personal preference. Reliance on the transcendent approach to quality has left many people in the translation industry with the impression that there is no way to measure the quality of a translation because a determination of appropriate accuracy and fluency ultimately derives from personal preference. However, I claim that the quality of translation can be measured. First, though, we must establish a more practical definition of translation quality.

In defining translation quality I rely heavily on the work of Alan K. Melby and other professionals who have contributed to various translation quality standards. The first step in formulating a definition of translation quality is to recognize that translation -- like many other human activities -- involves both a process and a product. The process of translation is the set of actions whereby a translation project is completed (this almost invariably involves the generation of a target text that corresponds to the source text). The product of the translation is the target

text itself. Sadly, the word translation is ambiguous in English and can mean either the process or

the product of translation. In this project, I have tried to use the word translation to refer to the

process of producing the target text rather than strictly the text itself. It is important that a

definition of quality apply to more than just the product for various reasons. For example, a

target text of a document used in legal proceedings may be perfectly written but be inadvertently

delivered to the wrong person resulting in an inappropriate disclosure of highly sensitive

information. Although the text was good, the party requesting the translation would rightfully be

dismayed at the poor quality of the process of completing the translation.

In order to address both process and product oriented aspects of quality, this project uses

the following specification-based definition:

> A quality translation achieves sufficient accuracy and fluency for the audience and
> purpose, while, in addition, meeting all other negotiated specifications that are
> appropriate to end-user needs.

SPECIFICATIONS include both tasks to perform and linguistic requirements for the target text, and

thus, this definition makes the implicit claim that translation quality includes aspects of both

process and product. This definition of quality is relative to aspects of the real world. There will

be some intended person or group of people who will read the translation. A good translation is

expected to meet the needs of those who use it. Accuracy and fluency in this definition are

constrained by the audience and purpose of the translation. Complete and faithful accuracy is not

necessary in all cases and may be a burden for some audiences and purposes when it conflicts

with fluency. Likewise, in some cases the factual information relevant to a particular audience

must be rendered accurately. Finally, this definition recognizes that the quality of the translation

is most important when the production of the translation is a business transaction, i.e. when some

party requests translation services from another party. The specifications that need to be met are

the ones negotiated by the parties involved; immeasureable, unspoken ideas or expectations about the translation should not factor into quality because they may not be shared by all of the parties or be appropriate for end-users.

With this definition of translation quality, it now becomes plausible to measure more than just the subjective value of the target text but the quality of the entire translation project. The definition proposed here works in tandem with the recent document ISO/TS 11669 (General Guidance -- Translation Projects). The framework for structured translation specifications provided in that document provides a consistent way for categorizing the specifications of a translation project so that the parties can agree on the specifications, know what they have agreed to, and so that the work of the translator—or in the case of this project, the post-editor—can be evaluated in respect to the specifications. There is no reason to expect the work of two or more translators/post-editors to be comparable unless they were given the same specifications. When we recognize the need for the above definition of translation quality, then the purpose of this project becomes clear, namely to provide a means of using specifications to evaluate the performance of a post-editor, which performance is viewed in terms of translation quality relative to project specifications.

2.2 PREVIOUS WORK. Up until the last ten years, very little research had been done on the subject of post-editing (Allen 2003). However, advances in MT have prompted an increase of interest in the subject (Allen & Hogan 2000; Alves 2003; Guerberof 2009; O'Brien 2002; O'Brien 2005; O'Brien 2011; Ramos 2010; Rios et al. 2011; Specia et al. 2011; Specia et al. 2009a; Specia et al. 2009b; Vieira & Specia 2011). That is not to say that this subject was unexplored during the twentieth century (Krings 2001; Veale & Way 1997; Wagner 1983), but rather that the last ten years have seen a substantial increase of research exploring this topic. Most previous studies

have focused on post-editing effort and the quality of the raw MT target text. This post-editing effort may be defined in a number of ways; most notably the work of Krings (2001) divides post-editing effort into three categories: temporal, technical, and cognitive. Temporal effort measures how long it takes the post-editor to finish editing the target text, whereas technical effort measures the changes made to the MT-generated text. The cognitive load is difficult to measure because techniques designed to measure the thought processes of translators/post-editors often make the task of translating more difficult (O'Brien 2005). However, O'Brien has found that a measure of cognitive effort can be obtained from other measures, such as comparing the differences between the changes of multiple post-editors and accepting pauses in the timed record of changes as an indication of increased cognitive activity (2005). But trying to guess the cognitive burden of post-editing may be unnecessary for determining the difficulty of a post-editing task because it has been shown that time alone is a reliable predictor of other effort measures (Sousa et al. 2011; Specia 2011). Specia has also developed a system of measuring expected post-editing effort so that companies can estimate whether a particular machine translated text is worth sending to post-editors (Specia & Farzindar 2010). However, the reliance on the transcendent view of quality poses a problem for studying post-editing effort.

Measuring the effort—or the time it takes—to post-edit a text assumes that the post-edited target text has sufficient quality in all cases compared and that the post-editor followed all of the procedures necessary for the project. The transcendent view of quality assumes that any translation is necessarily of the same quality as any other: the target text thereof has necessary accuracy and fluency. Measuring "the time it takes to post-edit a text" assumes that there is a definite endpoint for a translation and that there is only one way to approach translating any text (Melby et al. 2005). Once we recognize that translation quality is relative, measuring post-editing

effort is only useful when the specifications are the same. One machine translated text may be useless for a particular set of specifications while being suitable for another set of specifications. The amount of effort necessary to successfully post-edit a text in accordance with a set of specifications will probably change when the specifications change, even if the source, raw MT text, and post-editor are the same. Hence any measure of the usefulness of a particular raw MT text is only valuable when that measure is relative to a particular set of specifications; otherwise, the measure may be totally off base for the project at hand.

The approach to measuring the quality of post-editing espoused in this project solves the problems with the transcendent view of quality. This project provides a way for organizing and processing the information necessary to achieve an understanding of what factors are relevant to a post-editing project beyond temporal post-editing effort. Without a system for describing the requirements of a post-editing project, research concerning post-editing effort, especially temporal effort, cannot be reliably portable to real world projects. For example, a study may investigate how much time it takes to post-edit raw MT output into documentation for internal use in a software company. The results of this research would be difficult to relate to a project where the goal is to render a text for general public consumption. If the specifications are not explicitly stated, then the results of the study may be misinterpreted to be directly relevant to the subsequent project.

Moreover if the specifications are stated but not organized, a comparison of how relevant the research may be to the project at hand would be more difficult. If the above research example concluded that post-editing needed to take less than 10 minutes to be cost-effective, then such a measure might discriminate against good post-editors who take 20 minutes to post-edit a text in the second project. The reason for the difference in time may have less to do with individual

post-editors than it does with the project specifications. In other words, it takes more effort to post-edit a text for a general audience than it does for a small audience that is already aware of more of the context of the text. If project specifications remain implied, systems for assessing post-editing immediately become inaccurate. This is not to say that studies of post-editing effort that rely on the transcendent definition quality are not useful, but it is important to try to gain an understanding of what specifications were implicitly considered in the study and that future work involve explicit specifications.

Recently, Colina has proposed a rubric for assessing translation quality that corresponds well with the definition of quality used in this project (Colina 2008). Colina's rubric focuses on the product of the translation and takes the important step of recognizing that the audience and purpose of the translation are important. Moreover, Colina argues that it is possible, and expected, that the values for achieving certain categories in her rubric may be adjusted based on the circumstances of the translation. This variability in the point scale of the rubric is a central feature of this project as well. Angelelli has also proposed a rubric for assessing translator skill, which can also be taken as a measure of quality (Angelelli 2009). Angelelli's corresponds well with Colina's proposal (Hague et al. 2011), but this project has taken a direction more in line with Colina's work. In particular, examples for writing target text specifications were based largely on Conlina's rubric, but process-oriented specifications are particular to this study.

3. PROJECT DESIGN. A key component of this project is a format for describing structured translation specifications in a machine-readable fashion and a rubric derived from those specifications. An implementation of the format for describing structured translation specifications can be accomplished using the YAML data serialization language (Ben-Kiki et al. 2005). This project also includes software with a graphical user interface (GUI) that can handle specification creation, saving, and editing. I also wrote two GUI applications to aid in using the specifications, in the form of a machine readable rubric, as a tool for evaluating post-editing projects. This chapter will discuss the format for structured translation specifications, the rubric format, and the GUI programs, which are written in the Java programming language.

3.1 STRUCTURED TRANSLATION SPECIFICATIONS. Structured translation specifications are based on a predefined set of 21 translation PARAMETERS. A parameter is a heading for requirements that pertain to a translation project. For example, translation tasks require a particular language to translate the document into, called the target language. In the framework for developing structured translation specifications, which can be found at ttt.org/specs, this target language requirement is represented by the Target Language parameter. A SPECIFICATION is a value for a particular parameter. The specifications for a translation project are the values for the various parameters that represent the translation project's requirements.

In the framework for structured translation specifications provided in ISO/TS 11669, some parameters describe instructions to the translator or post-editor; others detail information important for understanding the translation task at hand (e.g. some parameters are used to describe the source text). Still other parameters pertain to various people who may work on the translation. In other words, the framework for structured translation specifications helps to describe the entire translation project beyond the requirements for the target text. The approach

of describing the entire translation project makes structured translation specifications useful for other purposes beyond post-editing assessment. For example, the Linport project incorporates specifications into a format for packaging translation materials (Linport 2012). In the Linport project, users may provide values to some or all of the 21 parameters in order to create a structured translation specifications document (generally formatted in XML). This document can be stored and passed along in a translation package.

This project uses the framework for structured translation specifications in a manner similar to Linport. However, my focus is evaluating post-editing rather than the transmission of materials. At this point, Linport formats specifications in XML as largely prose text. This allows for a great deal of flexibility in what values users can apply to various parameters. Unfortunately, it complicates computerized evaluation because a computer would need to parse natural language specifications provided in prose. This is not a problem for a system that is designed to package information where the expectation is that some human will eventually read and understand the textual content. This project attempts to aid people in using specifications to assess the performance of a post-editor by formalizing the specifications further than simply a organized list of 21 parameters.

3.2 FORMALIZED STRUCTURED TRANSLATION SPECIFICATIONS. This project proposes a format for formalizing structured translation specifications in order to support post-editing assessment. These formalized structured translation specifications (FSTS) naturally support the generation of a rubric for evaluating post-editing that can handle a high degree of variability in the specifications of various projects. The majority of components and types proposed in the FSTS format are shared with other systems of describing specifications, such as the hierarchy described in ISO/TS 11669 (General Guidance -- Translation Projects) and the status descriptors in the

Linport STS format (Linport 2012; Melby et al. 2011). Considering that 21 parameters exist and that most of the structure of the specifications has been described elsewhere, this report will only detail those major differences between FSTS and other specifications formats. A complete example of a valid FSTS document is provided in Appendix A with annotations. Examples of possible values for all 21 parameters can be found at ttt.org/specs.

This project employs the YAML 1.1 serialization language to represent FSTS data and the rubric. Using YAML is not a requirement to use the FSTS format, but a YAML parser is required to handle the FSTS documents created by software in this project. I used the SnakeYaml 1.9 library to write a YAML parser for FSTS data (2012). Software designed to use the FSTS format may do so by including a YAML parser that can handle the types described in this document which include all 21 parameters, DIRECTIVES, and a type for managing word COUNTS. However, wherever possible interested parties should use the parser written for this project. If developers choose to implement an FSTS format via XML, a relational database, or other medium, they should make sure to retain all the features described in this document, including all 21 parameters (as described), directives and the ability to link and merge specifications.

FSTS data consists of a hierarchy of CATEGORIES, PARAMETERS, ATTRIBUTES, and VALUES. The hierarchy described in ISO/TS 11669 (General Guidance -- Translation Projects) consists of four top-level categories that help to organize the 21 parameters: Linguistic, Production, Environment, and Relationships. The first category, Linguistic, contains two subcategories for parameters pertaining to the source and target text, Source and Target respectively. However, subcategories were not used in the FSTS format in this project for two reasons: first, only the Linguistic category is subcategorized; second, implementing the hierarchy

in Java is simpler if all categories have the same type of content. Instead, both subcategories were promoted to top-level categories with the following rationale.

The division of the Linguistic category into Source and Target subcategories is logical and effective because it follows the natural division between source and target texts already present in most translation tasks. If the Linguistic category were to be treated as a whole, it would include over half of all the parameters in the framework, 13 in total. This makes managing the Linguistic category as a whole difficult. For example, if someone were to display the four top-level categories as webpages, the Linguistic category would generally take up more than the average screen size resulting in a need to scroll the webpage, whereas the other categories— consisting as they do of on average three parameters—may often be displayed without the need to scroll. Even though the Linguistic category does not balance well with the remaining top-level categories, I would have retained it in the FSTS if it were not for the fact the doing so would significantly increase the complexity of the Java implementation.

It would probably be possible in Java to use polymorphism to allow for the value of a category to be either a set of parameters or a set or categories, but propagating such an object hierarchy to handle the entire FSTS would be difficult because the specifications of various parameters vary so widely (compare the Complexity parameter and the Typical Tasks parameter in Appendix A). A simple and elegant solution is to abandon Linguistic as a category and instead introduce Source and Target as top-level categories. This change helps to balance the various categories and add more consistency to the hierarchy; also, this is the only fundamental change to the overall hierarchy in respect to ISO/TS 11669 (General Guidance -- Translation Projects).

Therefore the top-level categories in the FSTS format are Source, Target, Production, Environment and Relationships. These five categories arrange 21 parameters into logical groups. Parameters contain attributes, which in turn have various values. Parameters that include a content attribute correspond to parameters that have no subparameters in other formats (see ttt.org/specs). The idea of a subparameter is rejected in the FSTS format. In an FSTS a parameter is a type and not a category, or label, meaning that the following analogy should hold for specifications: a parameter is to a specification as a class definition is to an object instance in object-oriented programming. If we treat parameters as categories, we lose the ability to constrain a particular parameter to hold only a particular set of content. Treating parameters as types helps to solidify the structure of an FSTS. Those parameters that take specifications that are not subdivided into various parts (such as language and region) have a content attribute, whereas those parameters that, in other formats, contain subparameters have the attributes necessary to organize their various contents.

All parameters have two attributes that assist in determining the importance of a particular specification for a particular project: status and priority. The value of the priority attribute is always an integer, and priorities will be discussed more in depth in following section concerning the rubric. The value of the status attribute is one of four options: INCOMPLETE, NOT SPECIFIED, PROPOSED, and APPROVED. An "incomplete" status means that the person initially writing the specification has not finished completing the specification. The default for a specification is "incomplete." In order to maximize the effectiveness of structured translation specifications, no parameters in an FSTS should have the status of "incomplete" at the time the translator or post-editor begins working on the target text. If a specification is to be left blank, meaning that it is not relevant to the current project, the status should be changed to "not

specified." If a specification contains some information but not necessarily all the information relevant to that particular parameter, or if a specification requires the input or approval of another party, the specification should have a status of "proposed". Once a specification has been approved, or if changing the specification is no longer an option (such as if compensation is non-negotiable), the status should be set to "approved." The status attribute is important because it indicates whether a specification has been sufficiently determined to proceed with the project; one cannot expect full compliance with specifications that are not approved. Other attributes are specific to individual parameters; see Appendix A for a complete list with examples.

With 21 parameters, some with upwards of seven attributes, a complete set of specifications is rather complex. Moreover, projects often share some specifications. For example, a document may need to be translated into several different languages, but in each case the Source specifications are identical. Or, two projects might require the same production steps for similar but different documents. Writing a complete set of specifications every time would be burdensome, so the FSTS format supports linking and merging of specifications. The Linport project has developed a system of storing specifications in MODELS that can be reused in later projects. This is achieved through a database that stores models of various specifications; whole categories of specifications can also be mixed to create a new complete set of specifications. However, there is no way to view models outside of the Linport project web page or to transfer a library of models to another application. Once the Linport specifications are saved to a file, any reference to what models may have been used to create the file are lost. The FSTS format is designed to allow for decoupling specifications even after serialization. Decoupling enables changing a whole category of specifications by simply changing the link. Without linking, it would be necessary to replace each specification by hand. Exactly how linking and merging are

implemented will depend on the language and system used to represent specifications, but this project provides one possible implementation using the YAML serialization language and an FSTS parser. A detailed example of linked specifications is provided in Appendix B.

One of the key components of this project is the use of DIRECTIVES to break down arbitrary prose descriptions into instructions that can be evaluated by a grader. Several parameters and attributes in an FSTS take a list of directives as their value. A directive is a single injunction to some member of the translation workflow in regards to the translation project. Since the focus of this research is evaluating post-editing, directives were designed with the intention of being used as instructions for the post-editor. A directive has two attributes: a request and a priority. The priority indicates how important it is that the request be fulfilled. The request consists of natural language content describing the post-editor's task. Figure 2 gives an example of a directive.

```
!Directive
priority: 10
request: The post-editor must change words and phrases that violate
         audience, purpose, or content correspondence requirements.
```

Figure 1. Example Directive

Some directives are written by the person who is creating or modifying the specifications; other directives are implied by providing values to certain parameters or attributes. For example, providing an Audience specification implies a directive that the post-editor or translation must make sure the target text is appropriate for the audience provided. Exactly what is appropriate for a particular audience is a theoretical question beyond the scope of this project. In any case, the

assumption behind the implied directive is that the person receiving the instructions will understand what is and is not appropriate for the audience specified.

A key question for this project is why some values are directives and others are not. I assert that the parameters and attributes most relevant to assessing the work of a single post-editor are the ones that require directives as their values. Future work may expand the use of directives, but for now the project is limited to identifying which specifications can be effectively broken down into directives to aid the assessment of post-editing.

3.3 RUBRIC. The rubric is generated based on FSTS data and relies heavily on the use of directives. In fact, in order to provide a structured system for evaluating almost all post-editing projects, the rubric is actually simply a list of directives in four categories: Target, Production, Environment, Relationships. When the person writing the specifications creates a directive, they have the option of specifying its priority. Likewise, when providing specifications, each parameter has the priority attribute, which may be provided to indicate the importance of a particular specification. In the GUI applications, the exact priority calculations are handled automatically, but in principle they should match the following description. Any parameter with a priority of zero and at least one directive should have its priority adjusted to match the sum of the priorities of its directives. If a directive has a priority of zero and it is assigned to a parameter that has a given priority, the priority of the directive will be priority of the parameter divided by the number of directives assigned to that parameter. Subsequently, the priority, or total possible points, for a given category is the sum of the priorities of the parameters contained in that category. This counting and adjusting is performed in a top-down then bottom-up fashion by the rubric software, although the calculations could be achieved by hand. This means that if a parameter has a priority of 10 and two directives, one with a priority of 10 and the other with a

priority of zero, the directive with a zero priority will be assigned a priority of 5 by the parameter (10/2=5). Then the priority of the parameter will be revaluated to equal 15 (10+5=15), and assuming that the two other parameters in the same category have priorities of 20 each, the total points possible for the category will be 55 (20+20+15=55). This algorithm for determining the priorities of directives, parameters, and categories is designed to make directive parameters primary while allowing for parameter-level management of priorities. In other words, individual directives are the most important, but users may use parameter-level priority estimates to avoid needing to micromanage individual directive priorities.

When using the rubric, an evaluator simply specifies whether a particular directive was fulfilled or not. If it was fulfilled, the value of the priority is awarded; otherwise, no points are awarded. The final score for a given category is the number of points received divided by the number of points possible. The total score for the entire rubric is the sum of the points received divided by the sum of points possible based on the priorities of the various directives. Even though this calculation results in a fraction, which is interpreted as a percentage, the format for the rubric retains all of the directives evaluated. Therefore, it is possible to review a rubric after having obtained a score to determine which directives a particular post-editor failed to achieve. Naturally, one cannot compare the work of post-editors who are given different directives because their scores would no longer be comparable. The rubric does not provide a way in and of itself to assess the competence of a post-editor; instead it provides a format for designing a test that may be used to assess performance.
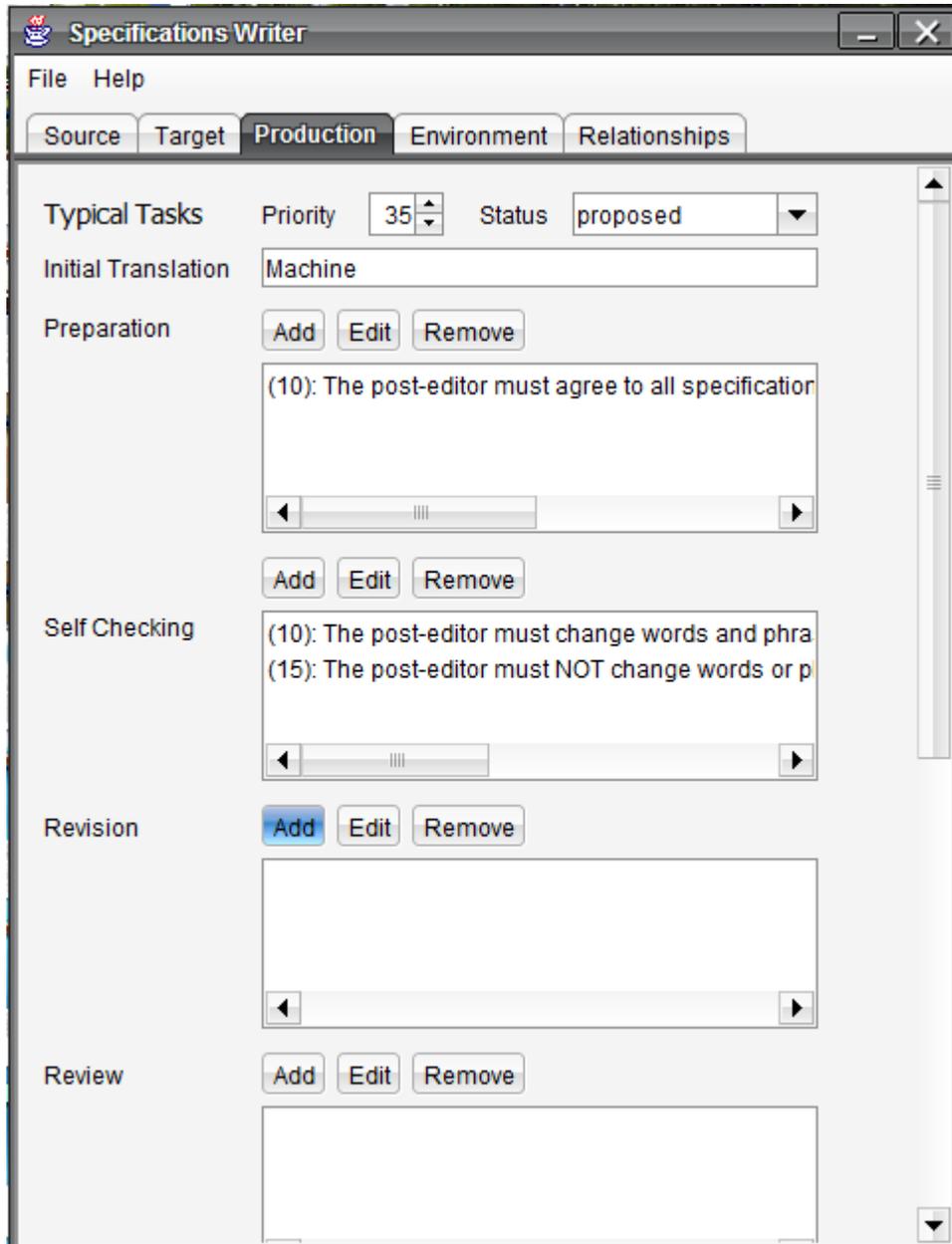
Figure 2. Specifications Writer Production Pane

3.4 JAVA IMPLEMENTATION. The software developed for this project is hosted as an open source

Google Code project at: http://code.google.com/p/ruqual/. The reader is invited to visit the

Ruqual project and download the software to explore while reading this section. The software

consists of two GUI applications: the Ruqual Rubric Viewer and the Ruqual Specifications

Writer. The Rubric Viewer is dependent on the same YAML parser as the Specifications Writer.

It is possible to load either an FSTS file or a previous rubric for grading. The interface for the

Specifications Writer consists of five tabs corresponding to the five top-level categories in the

FSTS as shown in Figure 3. There is a dropdown menu for assigning the current status of each

parameter and a widget for entering a number from 1 to 99 for the priority. Also, it is possible to

add, edit, and remove directives for those attributes that require a list of directives. The dialog for

creating or editing a directive is shown in in Figure 4. The Specifications writer allows for saving

and loading FSTS files so that it is possible to pause working on a set of specifications and return.

However, although loading linked and merged FSTS data is supported, the Specifications Writer

currently only saves FSTS data in a single file.
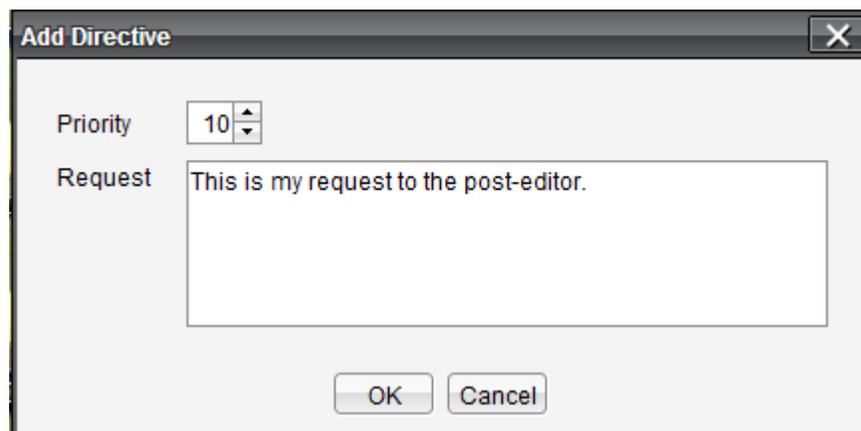
Figure 3. Specifications Writer Directive Dialog

The Rubric Viewer also has five tabs as well, but only four correspond to FSTS

categories. The reason is that the Rubric Viewer is designed to assist the user in evaluating the

work of a post-editor, which means that it displays all of the directives given to the post-editor

both explicit and implied. The Source category does not include any directives to the post-editor,

so it is not displayed in the Rubric Viewer. If the user needs to reference the complete set of

specifications, they can do so via the Specifications writer or by opening the FSTS file in a text editor. The fifth tab in the Rubric Viewer is the monitoring tab, shown in Figure 5. This tracks whether an answer of YES or NO has been provided for each directive in each category, as shown in Figure 6. If there is a directive that does not have an answer of YES or NO, the monitoring tab will display that category as "incomplete."
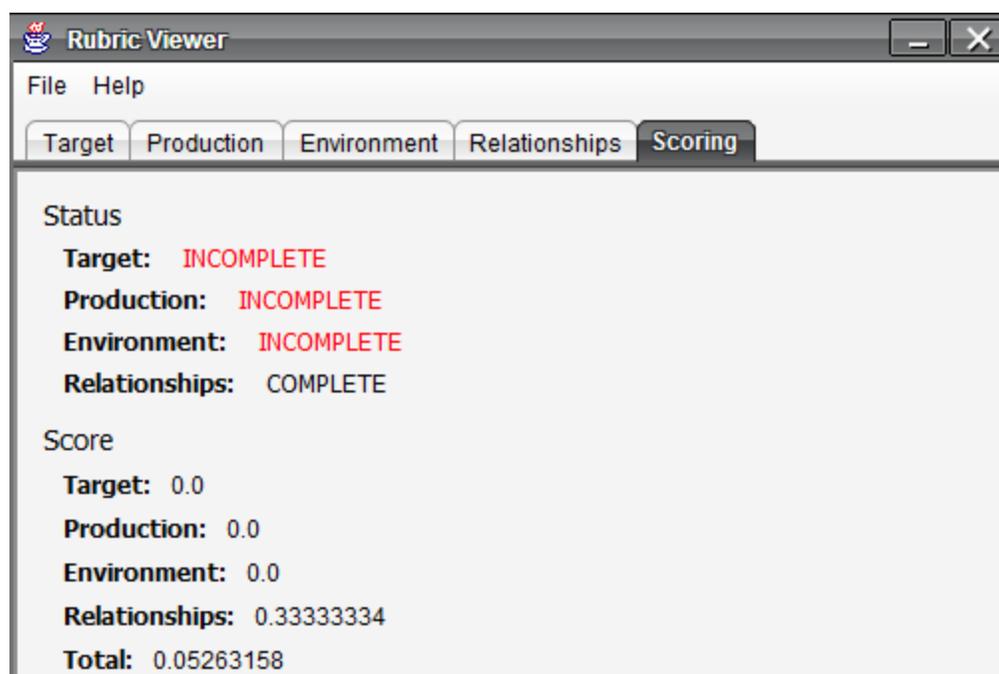
Figure 4. Rubric Viewer Monitoring Pane
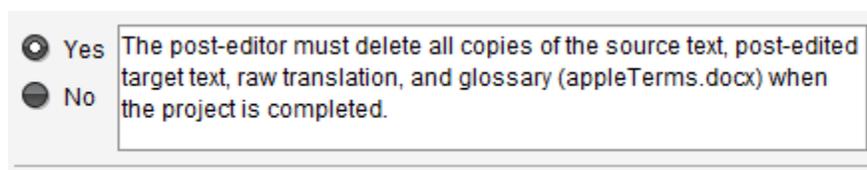
Figure 5. Example Directive Displayed in the Rubric Viewer

This software assists users in using the FSTS format and rubric without needing to be trained in reading YAML.

4. USER STUDY. One of the main benefits of this project is that it provides tools that allow users to assess whether the performance of a post-editor met the requirements of the translation project. The goal is to provide a means whereby people who are not highly trained in assessing the work of a post-editor can verify that a post-editor has done a good job on a particular project. As part of this project, I conducted a research study to examine whether non-expert graders agree generally on the quality of the work of five fictional post-editors. If the graders compared generally agree on the quality of the post-edited translations, then we would expect other non-expert graders to agree with the same determination. The claim is that with the tools provided in this project, it will be possible to write good specifications that allow non-experts to assess post-editing quality cost effectively. However, even though graders may agree on the quality score to assign to a particular post-editor, this score may not actually reflect what industry professionals would determine. In order to show that the scores provided by the non-experts are valid, an expert translator was also asked to assess the same post-editors. This separate expert grader was also asked to provide a strictly human translation in order to provide a comparison between the fictional post-editors and an actual translation. Although a variety of research may be conducted related to the FSTS format and this project's software, the user study is designed to answer the following research question in particular:

> How reliably can non-expert human graders assess the quality of post-edited machine translation, when the initial English target text was generated by a free and publically available machine translation system, when the source text is a medium difficulty (ILR level 2) document in Japanese, and when the graders assessing the performance of the post-editors are given a rubric based on a set of structured translation specifications?

This chapter will explain the design, participants, and implementation of a user study to answer the primary research question. Prior to surveying any human subjects, I obtained approval from

the BYU Institutional Review Board for Human Subjects by presenting the same methodology as described in this document (study number E120141).

4.1 DESIGN. The research question requires an ILR level 2 text as the source text for the study. The reason ILR level 2 was chosen is that non-expert graders, who have more limited Japanese skills than what we would expect from translation professionals, are expected to read and at least minimally understand the source text. The reason that they need to understand the source text is that they will be asked to discern whether there are any significant additions or omissions in meaning in the post-edited target texts. In other words, graders will need to be able to assess quality based on the definition given in Chapter 2, which includes questions of accuracy. The source text was obtained from a Japanese news website, Asahi Shimbun Digital (2009), and modified slightly to be more clearly ILR level 2. In particular, I conducted an analysis of cohesion in conjunction with a native speaker of Japanese, and she decided to add a short sentence to the end of the source text that introduces a major change in focus. The text has several null anaphora that already help to increase its complexity to ILR level two, while at the same time it is simple enough in terms of grammar and Kanji that non-experts should be able to understand it sufficiently. A complete copy of the source text is included in Appendix C.

The source text was translated by Google Translate (Google 2012) to produce the raw machine translated text, which can also be found in Appendix C. Joy Palmer also provided a human translated reference translation that was used in determining what alterations should be made to the raw machine translated text to create the five post-edited texts. Another key basis for the specifications used for this study and subsequently the alterations that provide the five post-edited texts is the work of several BYU language students who took the time to analyze the work of real post-editors. Actual post-editing project data was provided by Ray Flournoy of Adobe in

several different languages, and the students performed detailed annotations in German, Chinese, and Russian. The goal of the annotations was to identify what implied specifications the post-editors may have relied on to make their changes. Each student annotator was asked to fill out a complete set of structured translation specifications based on the changes they observed in the actual post-editing data. These data and annotations helped to influence the creation of specifications in two major ways. First, I was able to see what types of changes occurred in the post editing data. In particular, machine translated texts often need to be adjusted for syntax; and many times only a single word is replaced because the machine translated word is not appropriate for the situation, although it possibly grammatical in other cases. Second, I noticed that, at least in the case observed, alignment and processing are very important. It would seem that when post-editing is performed, it is usually performed in conjunction with other methods to facilitate rapid translation. This is one of the reasons why the directive to maintain sentence breaks was included in the specifications (see Appendix A: Content Correspondence).

In order to simulate post-editing, I modified the raw machine-translated text into five texts and included five scenarios describing the process the five fictional post-editors followed. These scenarios and texts are included in Appendix D. Some fictional post-editors violated more specifications than others in order to hopefully provide variation between the texts and scenarios. Also, errors were introduced—based on particular specifications—so that some fictional post-editors violated process-oriented directives while others violated product directives. Table 1 gives a complete overview of the fictional post-editors in terms of which parameters they were intended to violate. Post-editor D was engineered to violate the most specifications, whereas the post-editor C was intended to follow the most specifications. Letter identifiers were assigned randomly to try to eliminate any influence introduced by a grader recognizing a pattern such as A

is better than B and B is better than C, etc. The Target and Production specifications written for this study (see Appendix A) generally tend to address the PRODUCT of the translation (i.e., the post-edited target text), whereas the Environment and Relationships specifications address the PROCESS side of quality. This is not the only way to divide of the specifications used in this project. In fact those specifications in the Production category may actually describe steps to be taken in post-editing (i.e, process), but the tasks in the Production category deal primarily with the target text, whereas the Environment and Relationships specifications relate to other tasks or requirements that are not directly related to a particular text. In this respect, distinction between product and process in this document may be interpreted as a perspective on textual versus non-textual specifications.

Post-Editor	Specifications Violated	Specifications Followed
A	Production, Environment, Relationships	Target
B	Audience, Purpose, Target Language: region, Production	Environment, Relationships
C	Minor Production Violations, Target Language: terminology	All (generally)
D	All (generally)	Target (half)
E	Submissions, Permissions	Target, Production, Environment

Table 1. Fictional Post-editors

In order to try to make the post-editing scenarios and texts plausible, the middle post-editors were designed to generally violate either process-oriented specifications or product-oriented specifications rather than specifications selected at random.

4.2 PARTICIPANTS. In order to minimize costs, the translation industry is sometimes limited to using non-experts for evaluating purposes. For example, at a translation agency known to the author, the quality control agent was skilled in Chinese translation but was also required to do quality control of English translated from Japanese, Korean and other Asian languages. This

study examined whether someone with a limited knowledge of the source language may be able to render a reliable assessment. In particular, the definition of non-expert in this study was those individuals who are: non-native speakers of the source language, native speakers of the target language, at least high school graduates, completely inexperienced in the industry of translation although they may have experience performing translation for course work at the university level, and have at least two years experience studying the source language. For this study, those individuals who had been paid to render translation services were excluded from the definition of non-expert.

In order to determine whether non-expert graders were more than simply reliable with each other but actually able to match the judgments of translation professionals, an experienced Japanese to English translator was asked to take the same survey as the non-expert graders. This translator works as the project lead for Japanese translation at a successful translation agency. This translator contributed an evaluation of each of the five fictional post-editors with the same stimuli as the non-expert graders. Another translator provided a human translation of the source text, which helped to provide further understanding of how the specifications may be interpreted. The second translator has over 20 years experience translating Japanese in the area of patents and legal documents and also works as an in-house translator for a different company than the other expert translator.

In total 17 non-experts provided complete assessments of all five post-editors. The number of people who actually participated in the study was much higher, but not all participants met the definition of non-expert. Also, several people began taking the Qualtrics survey but did not finish. Over 20 people filled out almost complete translation rubrics, but 3 participants had to be thrown out because they seem to have missed some directives or demonstrated other

anomalies such as two or three identical rubrics. Of the 17 participants whose data could be counted, 5 were women and 15 were men. This bias towards men may be due to the fact that the survey was sent out to BYU students; a large percentage of Japanese bilinguals learned Japanese serving the Church of Jesus Christ of Latter-Day Saints on a mission in Japan. However, the survey was also sent out to non-BYU students such as the Idaho State University Japanese club and students at the Ohio State University. Therefore this difference between male and female Japanese bilinguals may be a larger trend. It may also be the case that the distribution of male and female graders is simply random chance. In all cases, gender is not of real concern in this study. On average the participants who completed the survey tended to be those who had studied Japanese for a longer period, generally over three years. The mean age of the participants was 24.3 years. This is likely due to the fact that older participants who have acquired Japanese may be more likely to have also worked in the translation industry, which would preclude them from being non-experts.

4.3 IMPLEMENTATION. The data were gathered via a Qualtrics survey that was accessible from the Ruqual website: ruqual.gevterm.net. The first portion of the survey asked some basic demographic questions, and then participants were directed to an instructions page that included four items:

1) A video demonstration of the Ruqual Rubric Viewer
2) A PDF text walkthrough with the same content as the video in case the participant lacked the software necessary to display the video
3) A place to download the source materials and terminology files
4) A link to a zipped version of the Ruqual Rubric Viewer

Participants were instructed to familiarize themselves with the software and source materials before proceeding with the survey. The next five questions presented each of the five fictional post-editors in random order. Each of these five questions included two links: one to the post-editor scenario and target text (see Appendix D), and one to a rubric file to be loaded and filled out in the Rubric Viewer. There was also a place to upload the completed rubric, which is how the rubric files were stored on byu.qualtrics.com. As compensation for participating in the study, those who completed the survey were given a pass code to be used on the Ruqual website whereby they were able to receive an Amazon Gift Card for 10 USD. Screenshots of the instructions page and an example post-editor question are included in Appendix E. The video instructions and walkthrough are available on request.

5. ANALYSIS. The analysis of the data is twofold. First, this chapter will detail a statistical analysis examining the question of reliability. As part of that analysis, I will present a descriptive analysis of the overall assessment of the five post editors by the non-expert graders. This will include a discussion of a possible ranking of the post-editors. Then, I will examine the question of reliability among non-expert graders, and finally I will turn to the question of whether non-experts are in concordance with an expert grader. Second, this chapter will present an analysis of the similarities and differences between the five fictional post-edited target texts and the real human translation provided by an expert grader.
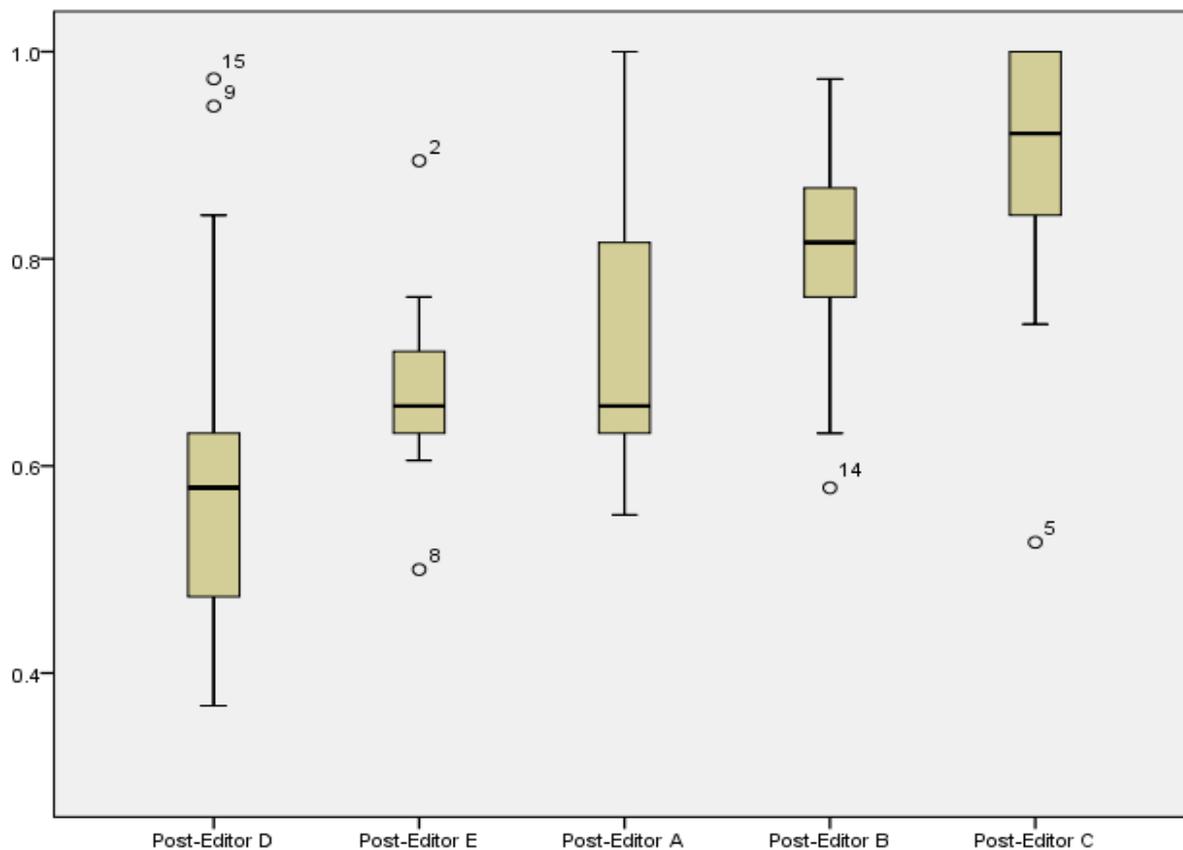
Figure 6. Post-Editor Scores Ordered by Median Score.

5.1 STATISTICAL ANALYSIS. The graders were instructed to assess each post-editor independently. In other words, the purpose of the study was not to create a ranking or comparison between post-editors. However, some post-editors were designed to violate more specifications and others less. Figure 6 shows a box plot of scores given to the five post-editors ordered by median score; this same ordering also holds for mean scores. On average, non-expert graders provided the highest score to post-editor C and the lowest score to post-editor D, which is what we would expect based on number of specifications these post-editors we designed to violate in Table 1. However, there were some outliers, such as graders 15 and 9, who chose to evaluate post-editor D very highly. These two graders also tended to rate other texts higher than average, which raises questions about how these graders chose to interpret directives. It may be the case that the instructions to graders have flaws that may account for these outliers. Assessments for post-editors E tended to cluster more tightly around the median, whereas there is more variation in scores for post-editors A, B, C, and D. It should be noted that no post-editor received a score lower than 0.35 from any grader, which could be due to the fact that all of the target texts were, in the opinion of the researcher, reasonably grammatical English. The Target Language: language attribute had a hard-coded default value of 50 for its priority because it is generally an important requirement that the target text be in the language specified. This means that each post-editor was expected to get a free 50 out of 190 points in all cases.

	A	B	C	D	E
Expert Scores	0.447	0.684	0.763	0.315	0.605
Upper 95% Confidence Limit	0.784	0.863	0.957	0.693	0.716
Lower 95% Confidence Limit	0.648	0.755	0.829	0.514	0.627

Table 2. Expert Grader Scores and Non-Expert Confidence Intervals

An expert grader also assessed post-editor C as the best and D as the worst, but the expert grader's scores do not directly match the ranking of the mean scores from the non-expert graders. The expert grader's assessments are provided in Table 2, and as demonstrated the expert grader provided a higher score for post-editor E and a lower score for post-editor A than the mean scores for those post-editors in the non-expert group. It should be noted that there is no statistical difference between post-editors A and E for the non-experts because the span of the "box" of post-editor E is within the span of the "box" of post-editor A; the data show that the mean scores do not match exactly the same order as the expert. There is a correlation between the experts scores and the mean non-expert scores for each post-editor, $r(3)=0.879$, $p < 0.05$, which suggest that the design goal of providing simulated post-editors with distinct differences and a progression from higher quality to lower quality may have been achieved. However, this correlation does not tell us whether non-expert graders are in concordance with the expert nor whether there is any kind of reliability among non-expert graders.

In order to test reliability, which is the focus of the research question put forward in this study, I calculated a two-way random Intraclass Correlation (ICC) along the lines of Shrout and Fleiss (Shrout & Fleiss 1979). ICC scores can range from 0 to 1 analogous to percentages. There are several advantages obtained from using an ICC, one of which is the ability to measure absolute agreement in addition to consistency. The question of reliability in this case is not simply whether graders assign the same relative scores to the post-editors but to what degree they are assigning the same absolute scores. Since all 17 non-expert graders assessed all five fictional post-editors and these graders are a sample of potential non-expert translation evaluators, the ICC calculated utilized a two-way random effects model with grader effects and measure

effects. Table 3 shows the single and average ICC scores for the non-experts graders subdivided by rubric categories.

Category	Single ICC	Average ICC
Target	0.167	0.773
Production	0.148	0.747
Environment	0.529	0.95
Relationships	0.607	0.963
Total	**0.426**	**0.927**

Table 3. Single and Average ICC Scores for Non-Expert Graders

The ICC is a measurement of the agreement between graders, or in other words, the percentage of variability in the scores that represents the quality of the post-editing. The single ICC is a measure of how reliable a single grader from this set of graders would be if we were to accept his or her score alone. The average ICC indicates at what percentage the graders agree with each other. The data in Table 3 are broken down into the four rubric categories and a category for the total score. The total is not the total of other ICC scores. The key statistic is the average ICC for the total score, which is ICC(2, 17) = 0.927. This is a strong indicator that the non-expert graders were reliable as a group. However, the single ICC for the same category was only ICC(2, 1) = 0.426 suggesting that if one grader were to be selected from this set, he or she would be expected to be reliable 43% of the time.

In terms of rubric categories, there appears to be a split between Target/Production specifications and Environment/Relationships specifications. This is worth noting because the specifications as constructed for the user study generally include product-oriented directives in Target/Production and process-oriented directives in Environment/Relationships. The graders might have an easier time agreeing on whether a post-editor followed the specified processes than they do deciding whether a particular text sufficiently corresponds with another text. Future

research is necessary to discover whether this distinction is the result of other factors in this study.

In addition to reliability there is also the question of whether non-expert graders are assessing the post-editors in a manner similar to how an expert would do so. I calculated a coefficient of concordance (see Appendix F) for each non-expert grader and the expert grader. Two graders showed a strong concordance with the expert, and 10 graders showed a moderate concordance. One grader showed a weak concordance, and 4 graders showed little or no concordance. The five graders who were outliers opposite to the trend in the relative scores for the post-editors in Figure 6, including graders 2, 9, 5, 14, and 15 are the same graders who showed weak or no concordance with the expert. This suggests that there may be some feature or features of the translation that both the expert graders and the majority of non-expert graders are using to determine quality. It also suggests that the quality scores being assigned may reflect the real quality of the post-editing when outliers are ignored. However, none of the expert grader's scores fall within the confidence interval of the non-experts mean scores, which provides evidence that there is a difference between the expert and the set of non-expert graders (see Table 2). The fact that the expert provided a lower score than the non-expert average in all cases indicates that the expert may be allowing the post-editor less leeway in the translation. The expert is expected to have a better understanding of the importance of following proper procedures and maintaining necessary accuracy in translation the expert. Therefore the expert is perhaps either more willing to find fault with the performance of the fictional post-editors or more aware of the failures present in the text and scenario. More research is necessary to what factors may influence this difference between experts and non-experts in terms of assessing post-editing.

5.2 EXPERT REFERENCE TRANSLATION DESCRIPTIVE ANALYSIS. A second expert provided a human translation without reference to the raw machine translated text. In fact, the second expert was only given the source text and specifications. This means that there should have been no influence from machine translation on the target text provided by this second expert. The purpose of requesting a human translation was to obtain a reference translation for the source text. None of the post-edited texts were intended to meet all of the specifications, but it would be worthwhile to identify how close these fictional post-edited texts might have been to full human translation via a comparative analysis. Since the graders generally selected post-editor C as having the highest quality, this analysis will focus largely on that text and whether it is reasonably similar to the human translated text given that both texts were intended to meet the same set of specifications. If post-editing is worthwhile and the human translation does in fact meet the specifications, then the post-edited text should be generally similar to the reference text. A copy of the human translated reference text is available in Appendix G.

Neither post-editor C nor the expert human translator followed the provided terminology exactly. The human translator did not use the term COMPANY STORE, and post-editor C did not use the term CEO, substituting OUTLETS and OWNER respectively. The human translator used the word COMMENTARY for the first line of the translation, whereas post-editor C used COMPANY OVERVIEW. This is a notable difference because the machine translation matched the human in this case, meaning that post-editor C should probably not have made a change in this case. In the third line of the translation, the human provided a longer form of the phrase "From PC to mobile phone" namely "From Personal Computers to Mobile Phones," which indicates that it may have been preferable for the post-editor to have corrected this from the machine translation. However, it is difficult to say for certain because perhaps the machine

translation was good enough for the specifications, in which case it was better for the post-editor to leave the translation as it was.

Another key difference between the post-edited versions and the human translation is the interpretation of the word など/nado/ and its scope. Based on the machine translated text, /nado/ was interpreted as having scope over its immediate nominal predecessor. This resulted in translations such as "mobile phones such as the iPhone" in the first sentence, whereas the expert human translator interpreted the sentence as a list of related objects to which /nado/ applies resulting in a broader scope and translations such as "the mobile phone iPhone, and other products." In my opinion, the expert translator's interpretation was more correct, and this shows how access to a machine translation may lead some post-editors astray when interpreting subtleties. In terms of the specifications, both such interpretations would probably be acceptable in this case, but the human translator's version certainly applies to the real world better because it does not suggest that Apple is in the business of selling different types of mobile phones other than the iPhone.

The human translator took advantage of the flexibility provided by specifications that allow for some awkward sentences provided that the target text fulfills its purpose for the target audience. The translator provided a couple of sentences that would be typically described as run-on sentences, but these sentences match the flow of the Japanese more exactly. In fact, such sentences may facilitate automatic alignment and processing better than the sentence breaks provided by the machine translation. Overall, it appears that post-editor C and the human translator are generally similar, but the requirement to not change sufficiently translated phrases in the initial machine translation limits the abilities of the post-editor. Perhaps post-editing can be used to meet a certain set of specifications, but the variation in the human reference

translation suggests that there may be cases where the use of an initial machine translation may

be too confining even when a high degree of accuracy or fluency is not required.

6. CONCLUSION. Overall, the results for the user study appear to support the hypothesis that non-expert graders can assess the quality of fictional post-edited translations at a high degree when taken as a group. This is a promising result for the project because it shows that it is possible to obtain agreement about the quality of post-editing when using formalized structured translation specifications. Moreover, the fact that a majority, 12 out of 17, of the non-expert graders showed at least a moderate concordance with an expert grader suggests that there is some similar sense of quality shared by the non-experts and the expert. Determining the source of the agreement was not the goal of the user study conducted. It may be that using the FSTS format and the Ruqual Rubric Viewer leads to increased reliability among evaluators, or perhaps the degree of agreement demonstrated in the user study is the result of other shared information. Without a control group or other method for controlling the influence of any shared knowledge independent of specifications, it is very difficult to say what role the specifications play in the reliability demonstrated. What is demonstrated is that graders do in fact agree.

The methodology of the user study did very little to control for a variety of possible confounding variables such as the environment in which the test was taken. Graders were permitted to complete the assessment in their own home or wherever they were able to obtain internet access. They were instructed to rely on the source materials provided and the specifications, but some may have used other resources such as dictionaries or the advice of friends or family. Furthermore, the very nature of translation makes providing a methodology that will consistently provide reliable and valid results very difficult. The use of specifications helps to allow a comparison of post-editors working on the same project; however, when the specifications change, so does the test. There is no guarantee that another study based on the same source materials and methodology will produce agreement when the specifications, and

hence the rubric, are changed. It is entirely possible to write misleading and unhelpful specifications, but it is the assertion of this project that the specifications used in testing are both practical and generally realistic for post-editing projects. Therefore, if practical specifications are provided and structured via this project's tools, then one would expect evaluators to reach reliable conclusions about the quality of post-editing.

This research suggests that it is possible to train post-editors to meet specifications and to also train evaluators to recognize the difference between good post-editors and bad post-editors in a manner that is consistent and replicable. Even beyond this purpose of producing reliable evaluators, there are other practical applications for this project. This project represents an authentic implementation of the ISO/TS 11669 standard. Using the tools helps to provide a consistent framework for describing translation, and in particular post-editing projects so that there is a greater degree of comparability in research of translation quality. Even though two tests examining translation quality use different specifications, comparisons may be possible between them based on the types of specifications used or the priorities set. Moreover, the FSTS format provides a format for describing and eventually transmitting the requirements of a translation project so that it should be easier to compare the requirements of two projects. The groundwork laid by this project should help in the development of other applications and the examination of other questions that take into consideration the specification approach to translation quality.

7. FUTURE WORK. This project provides a stepping stone for a broad variety of research projects relating to translation quality and the ISO/TS 11669 standard in particular. In the short term, several modifications to the software may be of immediate benefit. For example, people who access the current beta version of the software should be invited to provide feedback on the graphical user interface. A software design research question might be: how can the interface be improved to make it easier for translation professionals to understand and use? This project makes no claims that the current software implementation is the best software design for this problem; rather it is a first step in providing a tool for assessing post-editing. One potentially useful feature for users assessing post-editing is a window for doing a diff analysis of the raw machine-translated text and the post-edited text. In order to do this most effectively, it would be necessary to program the Ruqual Rubric Viewer to handle XLIFF files (XLIFF 2012). Pursuing an additional format other than FSTS for handling actual texts would not be advisable where open standards for such tasks already exist.

Another beneficial change to the implementation of this project would be localization. If translation professionals choose to use a version of the software as a standalone program, it would be ideal to translate the program into a number of major languages. This presents its own set of challenges because the terminology of the program should conform to ISO/TS 11669. Moreover, there is a question of whether to localize the actual FSTS format or if the categories, parameters, and attributes should remain in English to promote consistency. Initially, I would expect that maintaining English for the FSTS format would be more universal and acceptable, but future research might be done exploring the efficacy of one to one correspondences between languages for the relevant types in FSTS. Another key consideration for the FSTS format is an abstract formal definition of the format. This document provides a natural language description

of the format and required functionality, but a language neutral formal definition still needs to be addressed.

One natural application of this project is to move beyond the realm of post-editing. The principles used in this project should be applicable to not only the assessment of post-editing, but also to the assessment of target texts, translation quality assurance procedures independent of individual translations, cost estimation, the quality of an entire translation project from start to finish, the quality of a particular agency over a period of time, etc. However, the FSTS format and the software dependent on it were designed for post-editing. Some modifications to the format may need to be made to assess other aspects of translation quality. But I expect that the principle of using directives to break down instructions into units for evaluation will prove useful in all cases.

Another immediate application of this project is integration with the Linport project (Linport 2012), which relies on XML for formatting specifications. It should be possible with the tools currently available for XML to implement an FSTS in XML that can be used in Linport packages. It may also be possible to allow for YAML style FSTS data in Linport packages that may be converted as necessary to XML, depending on the processing applications requirements and the preferences of the developers. One advantage that might be obtained from an XML version is validation, but because not all specifications are necessarily required for all projects, validation is a lesser concern in my opinion. It is very easy to visualize specifications as a set of key value pairs, which is one of the reasons why this project promotes the YAML style FSTS.

Along with Linport integration, it may be more beneficial for this project to be implemented as a plug-in for a translation management system. The requirement to open a

separate stand alone program just for post-editing evaluation may be burdensome on quality control agents. In that light, the software for this project may serve as a proof of concept that translation tool vendors can use as a basis for supporting the FSTS format. Then the issue of an appropriate graphical user interface may be handled by each tool's own developers. Supporting specifications may become a major selling point for some tools especially if they also support the Linport format for transmitting translation materials.

Another major research question beyond reliability is the relationship between using an approach with structured translation specifications, such as FSTS, and evaluations based on less-structured instructions. People might generally agree on the quality of a translation because of a shared social definition of a quality translation, or people might not be as reliable as in this study without having structured translation specifications. This same question may also be examined from the expert versus non-expert axis to see whether general bilinguals share a different definition from translation professionals.

Likewise, the practicality of this project is limited by the fact that it assumes that some degree of proficiency in the source language is required. Naturally, if misinterpretation may lead to harm, then it is imperative that evaluators have the ability to verify the accuracy of the translation. However, there may be cases where fluency is paramount and confidence is placed in the translator/post-editor to make sure that sufficient accuracy is obtained. In such cases, it would be beneficial to investigate whether it is possible for evaluators who cannot understand the source text to provide assessments that show a concordance with those evaluators who are able to understand the source text. If an understanding of the source text is not a requirement for providing a realistic and reliable assessment when using FSTS, then this project may be able to provide a more cost effective solution than the research of this project currently supports.

A final key question raised by this project applies to the latter portion of the definition of quality, namely that a quality translation is one that meets specifications that in turn meet end user needs. This project only provides a method for deciding whether the performance of a post-editor meets the specifications provided. As discussed in the conclusion, it is entirely possible to write profoundly bad specifications that will not translate into quality for end users. More research needs to be done concerning what makes for good specifications and what tools can be developed to promote writing good specifications. This highly theoretical question will need to involve developing multiple sets of specifications, rather than the single set used in this project, and testing human translation consumers and professionals. Hopefully the work found in this project will help to add structure to such future research because a shared format is necessary in order to compare two sets of specifications accurately.

REFERENCES

2009. アップル.
   http://www.asahi.com/topics/%E3%82%A2%E3%83%83%E3%83%97%E3%83%AB.php.
   (Accessed on 2012).

2012. SnakeYaml. snakeyaml.org. (Accessed on 2012).

ALLEN, JEFFREY. 2003. Post-editing. *Computers and Translation: A translator's guide*, ed. by H. Somers, 297-317. Philadelphia: John Benjamins Publishing Company.

ALLEN, JEFFREY & CHRISTOPHER HOGAN. 2000. Toward the Development of a Post-Editing Module for Raw Machine Translation Output: A Controlled Language Perspective. Paper presented at the Third International Workshop on Controlled Language Applications, Seattle.

ALVES, F. 2003. *Triangulating Translation* Amsterdam: John Benjamins.

ANGELELLI, CLAUDIA V. 2009. Using a Rubric to Assess Translation Quality: Defining the Construct. *Testing and Assessment in Translation and Interpreting Studies: A Call for Dialogue between Research and Practice*, ed. by C.V. Angelelli & H.E. Jacobson, 13-47. Philadelphia: John Benjamins.

BEN-KIKI, OREN, CLARK EVANS & INGY DOT NET. 2005. YAML Ain't Markup Language Version 1.1. http://yaml.org/spec/1.1/current.html. (Accessed on 4/22, 2012).

BOOTH, ANDREW DONALD & KATHLEEN H. V. BOOTH. 1953. Automatic digital calculators. London: Butterworths Scientific Publications.

COLINA, SONIA. 2008. Translation Quality Evaluation: Empirical Evidence for a Functionalist Approach. *The Translator* 14.97-134.

GOOGLE. 2012. Google Translate. http://translate.google.com/. (Accessed on 2012).

GUERBEROF, ANA. 2009. Productivity and quality in mt post-editing. Paper presented at the MT Summit XII Workshop: Beyond Translation Memories: New Tools for Translators, Ottawa, Ontario, Canada.

HAGUE, DARYL R., ALAN K. MELBY & WANG ZHENG. 2011. Surveying Translation Quality Assessment: A Specification Approach. *The Interpreter and translator trainer* 5.243-63.

HUTCHINS, JOHN. 2007. Machine Translation: A concise history. *Computer aided translation: theory and practice*, ed. by C.S. Wai. Hong Kong: Chinese University of Hong Kong.

KRINGS, HANS P. (ed.) 2001. *Repairing Texts: Empirical Investigations of Machine Translation Post-editing Processes*. Kent, OH: Kent State University Press.

LINPORT. 2012. Linport: The Language Interoperability Portfolio Project. linport.org. (Accessed on 4/23,

MELBY, ALAN K. 2011. The Seoul of Standards and You. *The ATA Chronicle* 40.7-16.

MELBY, ALAN K., ARLE LOMMEL, NATHAN RASMUSSEN & JASON HOUSLEY. 2011. The Container Project. Paper presented at the First International Conference on Terminology, Languages, and Content Resources, Seoul, South Korea.

MELBY, ALAN K., ALAN D. MANNING & LETICIA KLEMETZ. 2005. Quality in Translation: A Lesson for the Study of Meaning. *Linguistics and the Human Sciences* 1.403-46.

O'BRIEN, SHARON. 2002. Teaching Post-Editing: A Proposal for Course Content. *6th EAMT Workshop Teaching Machine Translation*, Manchester, 2002.

—. 2005. Methodologies for Measuring the Correlations between Post-Editing Effort and Machine Translatability. *Machine Translation* 19.37-58.

—. 2011. Towards predicting post-editing productivity. *Machine Translation* 25.197-215.

RAMOS, LUCIANA. 2010. Post-Editing Free Machine Translation: From a Language Vendor's Perspective. . *Proceedings of the Ninth Conference of the Association for Machine Translation in the Americas (AMTA)*, Colorado, 2010.

RIOS, MIGUEL, WILKER AZIZ & LUCIA SPECIA. 2011. TINE: A Metric to Assess MT Adequacy. Paper presented at the 6th Workshop on Statistical Machine Translation (WMT-2011), Edinburgh.

SHROUT, PATRICK E. & JOSEPH L. FLEISS. 1979. Intraclass Correlations: Uses in Assessing Rater Reliability. *Psychological Bulletin* 86.420-28.

SOUSA, SHELIA C. M. DE, WILKER AZIZ & LUCIA SPECIA. 2011. Assessing the post-editing effort for automatic and semi-automatic translations of DVD subtitles. . Paper presented at the Recent Advances in Natural Language Processing Conference (RANLP-2011), Hissar, Bulgaria.

SPECIA, LUCIA. 2011. Exploiting Objective Annotations for Measuring Translation Post-editing Effort. Paper presented at the 15th Annual Conference of the European Association for Machine Translation, Leuven, Belgium.

SPECIA, LUCIA & ATEFEH FARZINDAR. 2010. Estimating Machine Translation Post-Editing Effort with HTER. Paper presented at the AMTA-2010 Workshop Bringing MT to the USER: MT Research and the Translation Industry, Denver, Colorado.

SPECIA, LUCIA, NAJEH HAJLAOUI, CATALINA HALLETT & WILKER AZIZ. 2011. Predicting Machine Translation Adequacy. Paper presented at the Machine Translation Summit XIII, Xiamen, China.

SPECIA, LUCIA, DHWAJ RAJ & MARCO TURCHI. 2009a. Machine translation evaluation versus quality estimation. *Machine Translation* 24.39-50.

SPECIA, LUCIA, CRAIG SAUNDERS & MARCO TURCHI. 2009b. Improving the Confidence of Machine Translation Quality Estimates. Paper presented at the MT Summit XII, Ottawa, Ontario, Canada.

VEALE, TONY & ANDY WAY. 1997. Gaijin: A Bootstrapping, Template-Driven Approach to Example-Based MT. Paper presented at the NeMNLP'97, Sofia, Bulgaria.

VIEIRA, LUCAS NUNES & LUCIA SPECIA. 2011. A Review of Machine Translation Tools from a Post-Editing Perspective. Paper presented at the 3rd Joint EM+/CNGL Workshop Bringing MT to the USER: Research Meets Translators (JEC 2011), Luxembourg.

WAGNER, ELIZABETH. 1983. Rapid post-editing of Systran. Paper presented at the Tools for the trade: Translating and the Computer 5, London.

XLIFF. 2012. www.oasis-open.org/committees/xliff/. (Accessed on 2012).

APPENDIX A: FSTS DATA

#A "#" indicates a comment.
#The symbol %% indicates places where an implied directive will be created.
#Statuses include "not specified", "incomplete", "proposed", and "approved".
#A ! indicates a type. All types other than !FSTS, !Count, and !Directive are #parameters in this document
#An FSTS is set of formalized structured translation specifications. This tag may be dropped when
#dealing with a parser that is aware of the root type.
Count is only used as the content of the Volume parameter. This basic count formalism may be
#replaced in the future.
#The priority attribute indicates the importance of the particular parameter or directive for this set of
#specifications.
#Empty lists are indicated by two braces as per YAML 1.1 (e.g. *[* and *]*, which may appear as *[]*).

```
!FSTS
specifications:
  #list of source parameters
  Source:
    textual characteristics: !TextualCharacteristics
      #picklist from iso 639-1
      language: ja
      #picklist from iso 3166-1
      region: JP
      #arbitrary text
      textType: Online Company Description
      #arbitrary text
      audience: General educated Japanese people
      #arbitrary text
      purpose: To inform people about Apple's products and history.
      #integer
      priority: 0
      #status picklist
      status: approved

    specialized language: !SpecializedLanguage
      #arbitrary text
      subjectField: Computers, Business
      #arbitrary text
      terminology: appleTerms.docx
      #integer
      priority: 0
      #status picklist
      status: approved

    volume: !Volume
      content: !Count
        #integer
        amount: 355
        #arbitrary text
        metric: Microsoft Word 2007 Word Count
        #arbitrary text
        unit: Japanese Characters
      #integer
      priority: 0
      #status picklist
```

status: *approved*

complexity: *!Complexity*
#arbitrary text
    content: *'ILR 2. The text is written to inform a general audience. There are some cohesive elements that may pose a chellenge to some translators and product names that require consistent translation. The translator will also need to be able to recover dropped verbs and manage common changes in tense and aspect.'*
#integer
priority: 0
#status picklist
status: *proposed*

origin: *!Origin*
#arbritrary text
content: *'The Asahi Shimbun Digital. A free online article at: http://www.asahi.com/topics/%E3%82%A2%E3%83%83%E3%83%97%E3%83%A B.php'*
#integer
priority: 0
#status picklist
status: *approved*

#list of target parameters
Target:
  target language: *!TargetLanguage*
  #picklist from iso 639-1, %%
  language: *en*
  #picklist from iso 3166-1, %%
  region: *US*
  #list of directives
  targetTerminology:
    - priority: 10
      request: *The target text must match terminology found in appleTerms.pdf.*
  #integer
  priority: 70
  #status picklist
  status: *approved*

  audience: *!Audience*
  #arbritrary text, %%
  content: *General Educated Americans*
  #integer
  priority: 10
  #status picklist
  status: *approved*

  purpose: *!Purpose*
  #arbritrary text, %%
  content: *To briefly inform people about Apple's products and history.*
  #integer
  priority: 15
  #status picklist
  status: *approved*

  content correspondence: *!ContentCorrespondence*

```
#list of directives
content:
  - priority: 5
    request: The text is to be adapted to the target language and region so that it does not generally
             appear to be a translation.
  - priority: 5
    request: There may be a few minor awkward expressions, but the text should still flow naturally in
             English.
  - priority: 10
          request: There may be minor alterations in meaning including additions and omissions
                   provided that the text still achieves its purpose for the target audience.
  - priority: 5
    request: 'The target text should match the overall complexity of the source text, which means that
             the translator should not introduce any technical terms or obscure references. '
  - priority: 5
     request: The target text should not break long Japanese sentences into smaller English sentences
              even if this results in a somewhat awkward sentence.
#integer
priority: 30
#status picklist
status: proposed

register: !Register
#list of directives
content:
  - priority: 5
    request: The target text should be written in a semi-formal style appropriate for mainstream news
             media.
#integer
priority: 10
#status picklist
status: approved

file format: !FileFormat
#arbitray text, %%
content: Microsoft Word 2007 or greater (.docx)
#integer
priority: 0
#status picklist
status: approved

style: !Style
#arbitray text
styleGuide: None Provided
#list of directives
styleRelevance: []
#integer
priority: 0
#status picklist
status: not specified

layout: !Layout
#list of directives
content: []
#integer
priority: 0
```

```
#status picklist
status: not specified

#list of production parameters
Production:
  typical tasks: !TypicalTasks
    #arbitrary text
    initialTranslation: Machine
    #list of directives
    preparation:
      - priority: 10
        request: The post-editor must agree to all specifications before beginning post-editing the raw
                 translation.
    #mapping of typical task names to lists of directives
    qa:
      #list of directives
      self-checking:  #post-editing – self-checking is the expected location for post-editing requirements
        - !Directive
          priority: 10
                  request: The post-editor must change words and phrases that violate audience, purpose,
                           or content correspondence requirements.
        - !Directive
          priority: 15
          request: The post-editor must NOT change words or phrases that are sufficiently translated in the
                   raw translation.
      #list of directives
      revision: []
      #list of directives
      review: []
      #list of directives
      final-formatting: []
      #list of directives
      proofreading: []
    #integer
    priority: 35
    #status picklist
    status: proposed

  additional tasks: !AdditionalTasks
    #list of directives
    content: []
    #integer
    priority: 0
    status: proposed

#list of environment parameters
Environment:
  technology: !Technology
    #list of directives
    content:
      - priority: 5
        request: The post-editor must have Adobe Acrobat Reader.
      - priority: 5
        request: The post-editor must use Microsoft Word 2007 (or greater) to edit the translation.
    #integer
    priority: 10
```

```
  #status picklist
  status: approved

reference materials: !ReferenceMaterials
  #list of references
  content: []
  #integer
  priority: 0
  #status picklist
  status: not specified

workplace requirements: !WorkplaceRequirements
  #list of directives
  content:
    - priority: 10
            request: The post-editor may not subcontract any portion of the work to a third party.
  #integer
  priority: 10
  #status picklist
  status: approved

#list of relationships parameters
Relationships:
  permissions: !Permissions
    #arbitrary text
    copyright: NA
    #arbitrary text
    recognition: NA
    #list of directives
    restrictions:
      - priority: 5
              request: The post-editor must delete all copies of the source text, post-edited target text, raw
                      translation, and glossary (appleTerms.docx) when the project is completed.
    #integer
    priority: 5
    #status picklist
    status: approved

  submissions: !Submissions
    #UTC timestamp, %%
    deadline: null
    #list of directives
    deliverables:
      - priority: 5
              request: The post-editor should return a copy of the post-edited text with the source text and
                      the raw machine translation.
    #list of directives
    delivery:
      - priority: 15
        request: The post-editor must email the deliverables before the deadline of March, 25 2012.
    #list of directives
    qualifications: []
    #integer
    priority: 50
    #status picklist
    status: approved
```

expectations: *!Expectations*
　#contact information and instructions
　communication:
　　contact: *!Contact*
　　　#should be either an email address or a mailing address
　　　address: *housleyjk@gmail.com*
　　　#arbitrary text
　　　name: *Jason Housley*
　　　#arbitrary text
　　　organization: *Brigham Young University*
　　#list of directives
　　instructions:
　　　- priority: 5
　　　　request: *The post-editor should confirm receipt of the source materials via email before starting the project.*
　#arbitrary text
　compensation: *Pro Bono*
　#integer
　priority: 5
　#status picklist
　status: *proposed*

APPENDIX B: LINKED SOURCE FSTS EXAMPLE

YAML supports merging of mappings with the merge indicator (<<), which makes it possible to copy a mapping and override various values. This appendix shows a snippet of an FSTS file that references another document containing source information. A link is indicated by a scalar beginning with the link indicator (->). The text immediately following the indicator must be a path to a file that is either relative to the file containing the link or an absolute path. Optionally, the filename may be followed by an @ which indicates that the contents returned should correspond to an anchor in the linked file (see the YAML specification for details about anchors and aliases (Ben-Kiki et al. 2005).) Therefore links should fit the following pattern: "->path/to/file@anchor". In the example, file named "ref.yml" contains only Source parameters. All of the Source parameters in the linked file are merged into the Source category. Then the Complexity parameter is overridden. All of the attributes of the Complexity parameter are copied from the linked file via the "complexity" anchor, and the content and status attributes are overridden. It is possible to copy the contents or a portion of the contents of one file either to merge the data with the merge indicator (<<) or to simply copy over the data such as for the status attribute below. It is actually possible to merge multiple mappings but mappings listed later will override previous mappings' keys. This system of linking and merging makes it possible to inherit various specifications for various types of projects. However, it is not possible to merge a list of directives. Lists and their members may be obtained via anchors but cannot be overridden because they lack an identifier to override.

```
!FSTS
specifications:
  Source:
    <<: ->ref.yml
    complexity: !Complexity
      <<: ->ref.yml@complexity
      content: ILR 3
      status: ->ref.yml@approved

#continued…

#Contents of a separated text file named ref.yml
#This file must be a in the same directory as the above FSTS for the link to
#be valid
textual characteristics: !TextualCharacteristics
  language: de
  region: US
  textType: Online Company Description
  audience: General educated German women living in the United States
  purpose: To sell makeup.
  priority: 0
  status: &approved approved
specialized language: !SpecializedLanguage
  subjectField: Popular Culture
  terminology: null
  priority: 0
  status: approved
volume: !Volume
```

```
  content: !Count
    amount: 10
    metric: Hand count
    unit: pages
  priority: 0
  status: approved
complexity: &complexity !Complexity
  content: ILR 2.
  priority: 0
  status: proposed
origin: !Origin
  content: ""
  priority: 0
  status: not specified
```

APPENDIX C: SOURCE MATERIALS

Instructions

This document includes the source text, a description of the source text, and the raw machine translated target text. These three texts were given to each post-editor. In addition, this document includes a list of definitions for common terms for your reference. Please review this document before downloading any post-edited target texts.

Source Text

解説

アップルとは

パソコンから携帯電話まで
マッキントッシュ（Ｍａｃｉｎｔｏｓｈ）などのパソコンや携帯音楽プレーヤー・アイポッド（ｉＰｏｄ）、携帯電話アイフォーン（ｉＰｈｏｎｅ）などを販売する、米コンピューターメーカー。世界中で直営店（アップルストア）やネット店を展開し、アイチューンズ（ｉＴｕｎｅｓ）ストアでは音楽や映画などの販売、アップストア（Ａｐｐ Ｓｔｏｒｅ）ではｉＰｈｏｎｅ向けのソフト販売も行う。世界市場ではマイクロソフトのウィンドウズで動くパソコンが９割以上を占め、同社のパソコンは数％のシェアしか得ていないが、斬新なデザインや使い勝手の良さなどでファンを獲得し独自路線を歩んできた。カリスマ的な経営者スティーブ・ジョブズ氏の言動が常に世界中で注目されることでも有名。しかし彼は２０１１年10月に死去した。

Source Text Description

textual characteristics:
    audience: General educated Japanese people
    language: Japanese
    purpose: To inform people about Apple's products and history.
    region: Japan
    textType: Online Company Description

specialized language:
    subjectField: Computers, Business
    terminology: see appleTerms.docx

volume:
    amount: 355
    metric: Microsoft Word 2007 Word Count
    unit: Japanese Characters

complexity: ILR 2. The text is written to inform a general audience. There are
some cohesive elements that may pose a challenge to some translators and product
names that require consistent translation. The translator will also need to
be able to recover dropped verbs and manage common changes in tense and aspect.
   origin: The Asahi Shimbun Digital. A free online article at: http://www.asahi.com/topics/アップル.php

Raw Machine Translated Target Text

Commentary
Apple and
From PC to mobile phone
Macintosh personal computer or portable music player such as iPod (Macintosh) (iPod), mobile phones to sell, such as iPhone (iPhone), the U.S. computer maker. Expand the net and shop (Apple Store), at (iTunes) store selling music and movies, also performs in software sales for the iPhone (App Store) Apple retail store in the world up. In the world market a personal computer running on Windows Microsoft accounted for more than 90%, PC company is not only get share of a few percent, has come a maverick won the fan, such as the difference between ease of use and innovative design. Well known for their words and deeds of Mr. Steve Jobs, the charismatic owner is always attention throughout the world. He died in October 2011, however.

Definitions

Target Text – This refers to the translated text returned to the requestor. The Raw Machine Translated Target Text is the output of a machine translation system, whereas the Post-Edited Target Text is the work of a post-editor who corrected the Raw Machine Translated Target Text.

Post-Editor – A post-editor is someone who corrects a machine translation to make it fit a particular set of specifications (audience, purpose, region, etc.)

Specifications – The specifications are the instructions and requirements given to the post-editor.

Requestor – The person or agency who asked the post-editor to participate in the project.

Project – The job of post-editing a particular text.

APPENDIX D: POST-EDITED TEXTS AND SCENARIOS

Post-Editor A

Scenario:
When the post-editor received the source materials, he/she sent them to an acquaintance who converted them to MS Word 2003 format because the post-editor did not own MS Word 2007 or greater. After finished post-editing, the post-editor sent the post-edited target text to the acquaintance to have it converted to .docx format, but this delayed the project meaning that the finished product was returned on March, 28 2012.

Target Text:
Company Description
Apple
From PC to mobile phone
Apple is the American computer maker that markets the Macintosh computer series, the iPod MP3 player, and the iPhone smart phone. Around the world, Apple has opened an internet store and company stores (Apple Stores). Music and movies are sold at the iTunes store whereas at the App Store software for the iPhone is sold. In the world market personal computers running on Microsoft Windows account for more than 90%. However, Apple only controls a few percentage points of market share, but Apple has taken its own approach to consumer electronics and won fans for its ease of use and innovative design. The company is well known for the words and deeds of Mr. Steve Jobs, the charismatic CEO, who was always earning the attention of the world. However, he passed away in October of 2011.

Post-Editor B

Scenario:
After getting the source materials, the post-editor verified via email that he/she had received all of the source materials and would follow all of the specifications. The post-editor owns and uses MS Word 2007 and Acrobat Reader. The post-editor completed and returned the post-edited text on March, 24 2012. He/she also included the source text and machine translation. Finally, the post-editor deleted all of the project files after hearing back that that project was complete.

Target Text:
Commentary
What's up with Apple?
From computers to mobile devices
Apple Inc. is a consumer electronics manufacturer that sells personal computers (Macs), iPod portable music players, and mobile phones such as the iPhone at pure bead prices. They also have retail stores (Apple Stores) and an internet store. In the iTunes store they charge extra for music and movies, and at the App Store they sell sub standard software for the iPhone. In the world market, machines running on Windows make up more than 90% of all machines. Apple

Inc., on the other hand, only controls a small percentage of the market. Apple Inc likes to do its own thing and has won some fans for its perceived ease of use and innovative design. Ultimately, the company gets a lot of attention for the words and deeds of Mr. Steve Jobs, the ostentatious owner, but he died in October of 2011.

Post-Editor C

Scenario:

Upon receipt of the source materials, the post-editor confirmed that he/she had received all of the source materials and would follow all of the specifications. The post-editor used Acrobat Reader to view the appleTerms.pdf file and MS Word 2010 to edit the target text. When finished post-editing, the post-editor returned the post-edited text alongside the source and machine translation on March, 23 2012. After completing the project, the post-editor deleted all of the related texts.

Target Text:

Company Overview
Apple
From PC to mobile phone
Apple is a U.S. computer maker that sells products such as the Macintosh personal computer, iPod portable music player, and mobile phones such as the iPhone. Apple has opened an internet store and company stores (Apple Stores) across the world. At the iTunes store they sell music and movies, while at the App Store they sell programs for the iPhone. In the world market personal computers running on Microsoft Windows account for more than 90% of the market, whereas Apple has only a small percentage of market share, but Apple has marched to the beat of its own drummer and won fans via ease of use and innovative design. The company is well known for the words and deeds of Mr. Steve Jobs, the charismatic owner, who is always getting attention throughout the world. However, he died in October of 2011.

Post-Editor D

Scenario:

The post-editor never agreed to the specifications. Instead, the post-editor returned the post-edited target text (by itself) on March, 26 2012 before a formal agreement was complete. After the project was complete, it was discovered that the post-editor had posted a copy of the source materials and his/her translation on his/her blog for the public to write comments about.

Target Text:

Commentary
And Apple?
From personal computers all the way to mobile devices:
Apple Inc. is in the business of consumer electronics. They make Macs, iPod portable music players, and the iPhone intelligent cell phone. They expanded their net shop and retail stores (Apple Stores); in the iTunes store they offer tunes and flicks, and at the App Store they have a

place for applications that run on the iPhone. In the world market a personal computer generally is running on Microsoft Windows, which accounts for more than 90%. Apple Inc. is a maverick to its fans, who love its innovative design and dang good usability. The company is well known for its attention getting owner Mr. Steve Jobs, but he died in October of 2011.

Post-Editor E

Scenario:
Because the post-editor did not own MS Word (and did not realize that he/she could download Acrobat Reader for free), he/she sent the source materials to a friend asking for help. Consequently, the friend was late returning the post-edited translation and did not include the original source text and raw machine translation. The post-editor hurriedly returned the finished post-edited translation on March 26, 2012, but he/she forgot to delete any of the project documents after the project was finished.

Target Text:
Company Overview
Apple
From PC to mobile phone
Apple is the U.S. computer maker that sells the Macintosh personal computer, iPod portable music player, and mobile phones such as the iPhone. They have company stores (Apple Stores) the world over and an internet store; at their iTunes store they sell music and movies, and at the App Store they sell software for the iPhone. In the world market, computers running Microsoft Windows account for more than 90%, while Apple has only a small percentage of the market, but Apple has followed its own path and won fans for its ease of use and innovative design. The company is well known for the words and deeds of Mr. Steve Jobs, the charismatic CEO, who is always getting attention throughout the world. He died in October of 2011, however.

APPENDIX E: QUALTRICS SURVEY DOCUMENTATION

Instructions Page:

You are going to assess the work of five post-editors. A post-editor is someone who corrects a machine translation (eg. Google Translate) to make the text fit a particular audience and purpose. You will need to determine whether a particular post-editor met specific requirements called specifications. In order to help you do this, you will use the Ruqual Rubric Viewer software.

Please begin by reviewing the Source Materials and Apple Terms. You do not need to translate or fully understand the source text (Japanese) in order to evaluate a post editor, instead you should make your best guess when trying to interpret the meaning or appropriateness of any of these texts. Your skills as a translator are NOT being evaluated, instead this study is focussed on your opinions of the post-editor's work.

It is important that you pay attention to the Machine Translated Target Text because the goal of post-editing is not to create a perfect translation but the most cost effective translation based on the machine translated version.

Please download the following files:
Source Materials
Apple Terms
Ruqual Rubric Viewer

Example Post-editor Question:

Post Editor A
Please download and review the following scenario file: Post Editor A
Please download the following rubric file: Rubric A

When you are finished evaluating the post-editor (when all of the areas on the scoring panel show: COMPLETE), please upload the saved rubric.

**If you have lost the Source Materials, appleTerms.pdf, or Ruqual Rubric Viewer, download this backup archive: Ruqual Backup

Choose File   No file chosen

>>

APPENDIX F: COEFFICIENT OF CONCORDANCE

Grader	PA-Total	PB-Total	PC-Total	PD-Total	PE-Total	Covariance	Variance	Mean	Coefficient of Concordance
1	0.632	0.816	0.868	0.392	0.632	0.032	0.035	0.668	0.817
2	0.921	0.816	0.816	0.842	0.895	-0.004	0.002	0.858	-0.063
3	0.658	0.842	0.737	0.474	0.605	0.020	0.019	0.663	0.657
4	0.684	0.789	1.000	0.605	0.658	0.024	0.024	0.747	0.517
5	1.000	0.737	0.526	0.474	0.737	-0.002	0.043	0.695	-0.040
6	0.658	0.895	1.000	0.579	0.632	0.029	0.034	0.753	0.564
7	0.553	0.763	0.921	0.526	0.605	0.027	0.028	0.674	0.744
8	0.737	0.684	0.842	0.368	0.500	0.024	0.036	0.626	0.657
9	0.895	0.868	0.947	0.947	0.763	-0.003	0.006	0.884	-0.040
10	0.605	0.947	0.816	0.526	0.658	0.026	0.029	0.711	0.632
11	0.658	0.868	0.842	0.526	0.684	0.024	0.020	0.716	0.632
12	0.816	0.816	1.000	0.632	0.763	0.020	0.018	0.805	0.368
13	0.868	0.895	0.974	0.684	0.711	0.016	0.015	0.826	0.269
14	0.579	0.579	0.921	0.605	0.632	0.016	0.021	0.663	0.501
15	0.658	0.974	0.974	0.974	0.684	0.006	0.028	0.853	0.085
16	0.632	0.842	1.000	0.632	0.605	0.025	0.030	0.742	0.518
17	0.632	0.632	1.000	0.474	0.658	0.028	0.038	0.679	0.680
Expert	0.447	0.684	0.763	0.316	0.605	0.033	0.033	0.563	1.000

APPENDIX G: HUMAN TRANSLATED REFERENCE TEXT

Commentary
Apple
From Personal Computers to Mobile Phones
Apple is a US computer manufacturer that sells Macintosh and other personal computers, the portable music player iPod, the mobile phone iPhone, and other products. Apple operates Apple Store outlets worldwide as well as an Internet shop, and sells music , movies, and other media at the iTunes store, and software for the iPhone in the Apps Store. Personal computers that operate Microsoft Windows account for over 90% of the world market, and Apple computers have only a small percentage of the market share, but due to their novel designs, ease of use, and other features, Apple computers have acquired fans and the company has walked an independent path. The words and actions of the charismatic CEO Steve Jobs were famous, being heard and seen worldwide. However, Steve Jobs passed away in October, 2011.