HUMAN AND MACHINE TRANSLATION QUALITY: DEFINABLE? ACHIEVABLE? DESIRABLE?

Alan K. Melby
Brigham Young University

Based on the presidential lecture delivered on August 10, 2012, at the occasion of LACUS XXXIX, York University, Toronto, Canada

THE DEBATE OVER MACHINE TRANSLATION (MT) began as the first general-purpose electronic computers were being built in the late 1940s. Warren Weaver (1955[1949]), then director of the Natural Sciences Division of the Rockefeller Foundation, wrote a highly influential memo suggesting that the use of computers in translation should be explored. Research projects aimed at programming those early computers to translate sprang up already in the 1950s. In the first period of work on machine translation, the goal was Fully-Automatic High Quality Translation (FAHQT), a term attributed to Yehoshua Bar-Hillel (1960), one of the first full-time researchers in machine translation.

During the Cold War, the desire to access research found in scientific journals published in Russia led to considerable US government funding for machine translation, especially Russian-to-English, during the 1950s and 1960s, until the scathing ALPAC (1966) report from the Automatic Language Processing Advisory Committee was published. The ALPAC report claimed that the quality of machine translation was very low and recommended a termination of funding for it.

Throughout the history of machine translation, in projects around the world, the question of quality has been constantly pursued yet ill-defined.

This paper was written in the context of a firm belief that although machine translation can be useful in some circumstances, it is not always an acceptable alternative to human translation. Professional translation is one of the most intellectually challenging of all human activities and will not be equalled by machines, unless machines eventually acquire full-blown artificial intelligence as currently seen only in sci-fi novels and movies.

After a long period of relatively little work on machine translation, at least in the United States, following the publication of the ALPAC report, development efforts revived in the 1980s, along with an alternative approach, namely, the use of computers as tools in the hands of human translators rather than as competition. Since then, a spectrum of human and machine involvement in translation has been in place, with the need for translation increasing because of factors such as increased global trade. Points on the spectrum from one extreme to another are sometimes labeled as shown in **Table 1** (overleaf).

In this paper, machine translation will be used to refer to fully automatic translation, regardless of its quality. The term *raw (unedited) machine translation* is used to emphasize that no human help has yet been enlisted to correct errors in the output of the system.

Abbreviation	*Explanation*
FAHQT	*Fully Automatic High-Quality Machine Translation* with no human involvement;
HAMT	*Human-Assisted Machine Translation*, i.e., MT plus post-editing and/or pre-editing;
MAHT	*Machine-Assisted Human Translation*, i.e., human uses computer for help, as desired;
HTLGI	*Human Translation* (unassisted by any technology) *"Like God Intended"*. This extreme was proposed tongue-in-cheek from the beginning in order to provide symmetry with the other extreme of machine translation unassisted by humans; few human translators today avoid all technology, typically using at least word processing, electronic dictionaries, email, and Internet search engines.

Table 1. *Spectrum of human and machine involvement in translation.*

Machine translation is at the center of Human-Assisted Machine Translation (HAMT). An entire text is translated automatically, and then, typically, a human post-editor fixes mistakes in the raw machine translation, until the text is acceptable for the purpose at hand.

Human translation is at the center of MAHT. The human translator has a variety of optional resources available, such as entire sentences from a 'Translation Memory' database of segments of text and their human translation, automatically retrieved when they are identical or similar to the current sentence being translated. Terminology is automatically looked up. And, in some cases, a machine translation of the current sentence is displayed, but it can be ignored at the option of the translator if it is of insufficient quality to edit.

The question of quality comes up repeatedly in discussions of machine translation. But what is translation quality? Translators often reject machine translation because of its low quality. Machine translation researchers sometimes claim that the quality of human translation is not so high. Yet there is no widely accepted definition of translation quality that applies to the entire spectrum of human and machine translation just described, and is used to measure quality reliably.

One purpose of this paper is to contribute to the debate about the role of humans and machines in translation by proposing and applying a new definition of translation quality.

The intended audience of the paper is broad, including:

1. academics who study language and are called linguists,
2. government language workers, who translate and who are also called linguists, and professional translators and translation project managers.

The structure of the rest of the paper is a sequence of three questions about translation quality and proposed answers to them.

1. Definable? Is translation quality definable? Not everyone thinks so, but I do. My definition builds on a standardized framework for developing structured translation specifications, which will be explained.
2. Achievable? Is quality achievable when machine translation is involved? This question will be answered in terms of the proposed definition of translation quality.
3. Desirable? Is translation quality desirable? Of course, quality human translation is desirable, but what about quality machine translation? Suppose it is desirable. Should everyone focus on FAHQT? Are there less challenging but still useful goals to be pursued while the ultimate machine translation system is under development? Again, this question will be answered in terms of the proposed definition by defining two translation-specific variations of the classic Turing Test.

In the conclusion of this paper, I will make some predictions concerning machine translation before the year 2045, but more importantly, I will suggest some short-term action items for those who are willing to give the new definition of translation quality a try.

1. DEFINABLE? Translation quality is notoriously difficult to define.[1] Consider the following quote from a European Commission document about the cost of poor translation quality in a practical translation environment:

> Although, like quality in general, quality in translation is a somewhat elusive concept, poor quality translations are in some—though not all—cases rather easy to detect. At best, a poor or less fortunate translation makes the reader shake his head

[1] For a survey of a number of mutually incompatible approaches to translation quality in the translation studies literature, I refer the reader to House (2001), even though I disagree with her characterization and quick dismissal of Skopos theory, as proposed by Reiss and Vermeer. House says one should examine three aspects of a theory of translation:

> [T]he nature of (1) the relationship between a source text and its translation, (2) the relationship between (features of) the text(s) and how they are perceived by human agents (author, translator, recipient), and (3) the consequences [that] views about these relationships have for determining the borders between a translation and the other textual operations (1997:1).

With respect to Skopos theory, she defines these aspects as follows:

> [T]he functionalistic [Skopos] approach is not concerned about the relationship between original and translation, nor is it concerned with establishing criteria for delimiting a translation from other textual operations. As it stands, functionalistic approaches are solely concerned with the relationship between (features of) the text and the human agents concerned with them (1997:16).

I disagree with her conclusions, at least for how Skopos has evolved as elaborated by Christiane Nord. Even if issues of correspondence are not in the foreground, functionalist approaches would argue that for the text to serve a desired function, some appropriate correspondence between source and target would have to pertain. As a result the issue of correspondence and its relation to other factors is omnipresent.

> and smile at a poorly translated sentence, but errors in translation can also have serious legal, financial or political consequences. (European Commission 2012:13)

We will later return to Functionalism in translation studies as a descendant of Skopos theory, to show how it helps address this question.

1.1. COVERT VS. OVERT TRANSLATION. An important distinction made by House is between covert and overt translation (2001:249–50, 2010:245–46). An overt translation, as the name implies, makes no attempt to hide the fact that it is a translation. According to House, in an overt translation the translator is asked to "give target culture members access to the original text and its cultural impact on source culture members". In contrast, a covert translation is an attempt to create a document that corresponds to the source text while appearing to have been authored in the target language and culture. As House points out, "The result may be a very real distance from the original". A commonly committed error encountered over the centuries in attempting to define translation is the tendency to privilege one type of translation, such as overt or covert, in the definition, as the only right way to translate. Can both overt and covert translations be considered quality products?

For example, an overt translation might choose to leave some culture-specific concepts untranslated, while a covert translation might make considerable adaptation. Although J.R.R. Tolkien's *Lord of the Rings* is not actually a translation (it is rather a literary "pseudo-translation", i.e., it portrays itself as a translation of an ancient text), Tolkien's appendices (particularly Appendix F, section II, "On Translation") explain how he adapted names and concepts from his putative ancient source in order to make the text approachable for a modern audience (i.e., making a covert translation). For example, he "Englished" many place and character names (i.e., created familiar-sounding equivalents for them), such as English *Shire* for *Sûza* and *Sam* as the equivalent of *Ban*, the short form of *Banazîr* in the fictional language of the Hobbits (Tolkien 1965[1955]:477–79). Even if *The Lord of the Rings* uses translation only as a literary device, these notes ring true to the issues translators do face when encountering names or concepts that are out of the world knowledge of the target audience. Translators often face the dilemma of what to translate into familiar terms and what to translate in unfamiliar terms because the concepts (rather than the words) involved may themselves actually be foreign.

In the case of a real translation (albeit a subtle one between two varieties of English), the U.S. publisher Scholastic changed the name of the first Harry Potter book from *Harry Potter and the Philosopher's Stone* to *Harry Potter and the Sorcerer's Stone* for the U.S. release and translated other cultural terms into U.S. English (e.g., *crumpet* to *English muffin* and *jumper* into *sweater*), a very covert sort of translation in the sense that most readers would never know the changes had taken place. The U.S. editor, Arthur Levine, was quoted as saying:

> I wasn't trying to, quote, "Americanize" them. What I was trying to do is translate, which I think is different. I wanted to make sure that an American child reading the books would have the same literary experience that a British kid would have. A kid

> should be confused or challenged when the author wants the kid to be confused or challenged and not because of a difference of language. (quoted in Radosh 1999)

An examination of various editions shows that it was not only American translators who made such adaptations: Jean-François Ménard, the French translator, for instance, invented names to try to make the text accessible for French readers, and thus "*Snape* became *Rogue*, *Slytherin* became *Serpentard*, and the British word *Bagman* became *Verpay*, from the acronym *VRP*, describing someone engaged in door-to-door sales" (Goldstein 2004, edited). Similarly, the Russian translation of *Lord of the Rings* renders "Strider" as *Бродяжник (Brodyazhnik)*, translating the meaning of the English name, while translations into some other languages leave the the name in its English form. (Robert Orr, pers. comm.)

Although the target audience for Harry Potter books seemed unfazed by these covert translations, not all readers were happy, leading to criticism, particularly of the American adaptations as cultural imperialism or "dumbing down" for an American audience (see, e.g., Gleick 2000). Given the difference in expectations of the target audience (children reading "juvenile literature" with little or no interest in the theory of translation and culture) and the professional translator critics, for whom such concerns are vital, it is difficult to imagine any translation on the overt-covert continuum that would have satisfied all readers.

At the other end, an overt translation might resort to an extensive apparatus of notes and other explanatory materials external to the translated text itself in order to render it accessible to readers. For example, the Chinese translation of James Joyce's Ulysses was a very overt translation and employed 5,991 footnotes (the greatest number of footnotes ever in a Chinese-language book) in an effort to explain the linguistic and cultural intricacies of Joyce's text to Chinese readers (Murphy 1995), perhaps leading to questions about the boundary between "translation" and "analysis".

Some scholars argue that the choice between overt and covert is not just a pragmatic one, but also an ethical one. Lawrence Venuti (1998, 2009), for example, describes the overt-covert distinction as the tension between a foreignized and a domesticated translation and has criticized Functionalism on several points. He maintains that part of the duty of the translator is to make the foreignness of the text apparent to readers, even at the expense of readability and accessibility. Venuti sees this position as an antidote to absolutist positions concerning translation that seek to domesticate it. A description of these criticisms and responses to them is found in Hague, Melby, and Wang (2011). Can both overt and covert translations be high quality? It depends on how one defines *translation* and *quality*.

1.2. DEFINING TRANSLATION. To the above evidence of the difficulty of defining translation quality, I add an anecdote about Roy Harris, who later became a respected professor of general linguistics at the University of Oxford, as told by David Bellos (2011:3): supposedly, early in his career at Oxford, Harris was assigned to teach translation but got out of it by refusing to teach something the faculty board could not define. Not only is translation quality difficult to define, but translation itself is.

Thus I take a small detour to define translation. First I will assume some building blocks, namely, source language and target language. One translates from the source language into

the target language. Source text and target text are the respective documents one begins with and creates. Even this much is controversial, for Roy Harris disputes the notion of a language (Harris 1982). Actually, there is much to dispute, and commercial translation companies recognize the need to distinguish among *locales*, which are geographic regions, such as the United States and the United Kingdom, together with their various conventions, such as for dates (is "8/12" the *12th of August* or the *8th of December*?), times, and currency. English does not exist as a well-defined object; witness the many differences between American and British English. In addition, O'Sullivan (2013:7) argues for a broader notion of "multimodal translation" that involves content beyond text, such as graphics and other non-linguistic content. Nevertheless, I will move on and use source/target *language* and *text* in the rest of this paper, using "text" and "content" interchangeably, without apology.

A welcome point of agreement in the world of translation is that the word translation exhibits a basic ambiguity between the *process* of translating and the *product* (i.e., the result) of translating. Below I will provide both process and product oriented definitions.

However, I part ways with most definitions of translation because they use the *transfer* metaphor, as in defining translation as the transfer of meaning from a source text to a target text. Elsewhere, my colleagues and I have discussed at length the notion of meaning as it relates to translation (Melby, Manning & Klemetz 2007) and have shown, we believe, that meaning does not reside in a text but rather is dynamically created from text and context in the mind of a human. If we are right, then defining translation as the transfer of meaning from one text to another is a dead end. Meaning is something we create when we interpret or create a text, but it never leaves our mind and thus never gets embedded in a piece of paper or other medium for representing a text. Therefore, meaning cannot be transferred from a source text to a target text. The translator assigns a meaning to the source text, hopefully similar to the intended meaning of the author, and creates a target text to which a reader will hopefully assign a meaning that also corresponds to the author's intent adjusted as required for the target audience.

Roman Jakobson (2000[1959]) suggests a definition that avoids claiming that meaning resides in a text. He defines translation as the act of interpreting a verbal sign, and he distinguishes among three types of translation: *intralingual* translation (commonly called paraphrasing or rewording); *interlingual* translation (between two languages); and *intersemiotic* translation (between a human language and some nonverbal sign system). Today, translation is assumed to be interlingual translation unless clearly indicated otherwise. For Jakobson, translation is an act of interpretation that assigns meaning to a text rather than extracting meaning from a text. The directionality of Jakobson's definition is thus the opposite of the directionality in traditional definitions of translation. (However, this argument is not one of unfettered relativism: the text itself certainly does have a profound influence on meaning, for the reader is always faced with a particular text rather than another. It would certainly stretch credibility, for example, to argue that a service manual for a 1964 Ford Mustang is really a treatise on Zen Buddhist meditation practices, despite the famous connection between Zen and motorcycles.)

Neither Jakobson's nor any other definition of translation can be fully theory-neutral. I have chosen a definition of translation that is compatible with Jakobson but goes beyond it by adding an element called specifications.

> Process definition: **Translation** is the process of creating target-language content that corresponds to the source content according to agreed-upon specifications.

Alternatively, one can take a product-oriented approach. Thus, a translation product is the result of some process involving humans or machines or both that results in a target text that corresponds to its source text according to agreed-upon specifications.

1.3. STRUCTURED TRANSLATION SPECIFICATIONS. Rather than depending on a definition of meaning, the above definition of translation depends on an elaboration of translation specifications to avoid becoming impossibly vague. Such an elaboration exists and is found in a document published by ISO, the International Organization for Standarization (www.iso.org), that provides guidance for translation projects: ISO/TS-11669 (2012).[2] As editor of this document, I worked with delegations from many countries over a period of five years, finally arriving at a substantial degree of consensus on defining 21 translation parameters whose values for a particular type of translation become a customized set of structured translation specifications (structured in that they keep the order of the parameters). These 21 parameters are divided into five main categories (source description, target requirements, process, project environment, and stakeholder relationships) under three broad aspects of translation (product, process, and project) and are available to the public with descriptions of each (Parameters 2013).

- **Product**
  - *Source description parameters* include those parameters that describe the source text, such as its language and region.
  - *Target requirements parameters* specify requirements for the target text and its relationship to the source text, such as content correspondence, which may ask for either an overt or a covert translation.
- **Process**
  - *Process parameters* address the tasks to be performed, such as initial translation, revision (a bilingual task involving a comparison of the source and target texts to determine whether they correspond according to the product specifications), and review (involving a subject-matter expert to determine whether the target text makes sense in a particular subject field, such as *medicine* or *physics*). The initial translation can be classic human, raw machine, or some combination.
- **Project**

[2] In addition, the ASTM F2575-14 standard on translation quality management has adopted the same set of parameters (see http://www.astm.org/Standards/F2575.htm).

- *Project environment parameters* address issues about the environment in which the project is to be carried out, such as whether the work must be done in a secure facility or whether specific technology must be used.
- *Project relationship parameters* address the relationship between the buyer and seller, including the fundamental issues of cost and delivery deadline.

There is nothing mysterious about any of these parameters, but they make the definition of translation at once precise and flexible. Not all quality translations are the same, but they do not vary chaotically. Each translation is assessed relative to a set of structured specifications derived from the universal framework of parameters.

Translations are typically assessed according to just one category of parameters, target-text requirements, in which case it doesn't matter to the assessment how long it takes to produce the target text or what steps were followed to arrive at it, or, alternatively, translations can be assessed according to multiple categories of parameters.

The mention of assessment indicates that a definition of translation quality is nearby, but in order to properly lay the foundation for that definition, I must introduce three types of translation stakeholders and five perspectives on quality in general.

1.4. TRANSLATION STAKEHOLDERS. Someone asks for a translation. Someone gets the translation done. And someone uses the translation. In ISO/TS-11669, these stakeholders are called the *requester*, the *provider*, and the *end-user*. Sometimes the requester and the end-user are the same person, such as when someone asks a colleague to translate a short passage while he or she waits, but typically the requester and the end users are different. The term *client* can be understood to be either the requester or the end user and therefore will be avoided for clarity and because it is ambiguous and implies a commercial arrangement; yet translations can be requested and provided within an organization without any money changing hands.

1.5. PERSPECTIVES ON QUALITY. There is a literature on quality in general that has been largely ignored in the world of translation. Garvin (1984) distinguishes five perspectives on quality that apply across many industries: (1) the transcendent approach, (2) the manufacturing-based approach (which we will call the 'production-based approach' in this article), (3) the user-based approach, (4) the product-based approach, and (5) the value-based approach.

1. The *transcendent* perspective assumes that there is an absolute ideal and that quality is measured by how close one comes to that ideal. In cooking, one might seek to bake the perfect apple pie. In music, an orchestra might strive toward the perfect performance of Beethoven's Ninth Symphony.
2. The *production-based* (*manufacturing)* perspective is focused on compliance with specifications. Quality consists in meeting specifications, whatever they are. An electric motor can be assembled using quality components from any of several sup-

pliers, and it will function equally well, since those components meet the specifications within extremely close tolerances.

3. The *user* perspective is focused on the end-user experience. Quality is measured by how well a product or service meets the needs of the end user. In a sense, the user perspective is about the adequacy of specifications. For example, suppose that a walking shoe has been manufactured according to the given specifications, but those specifications included a requirement that the front of the shoe extend an extra centimeter and point down 13 degrees. The shoe may look stylish (or strange) but the problem is that it is not good for walking. The extended toe will cause most people to trip unless they curl up their toes uncomfortably. The specifications were followed, but they were inappropriate. They were not "fit for purpose", since the purpose of the shoes was to support comfortable walking.
4. The *product-based* perspective quantifies the quality of a product or service based on ingredients or attributes. For example, in ice cream a higher butter fat content is considered a marker of higher quality, and for many types of cloth a higher number of "ends per inch" indicates a higher quality. This perspective generally favors empirically measurable and verifiable attributes over personal preferences, although some dimensions may be more subjective than others (e.g., an evaluator of the quality of cloth might be asked to rate the smoothness of its finish based on handling the cloth).
5. The *value-based* perspective assesses quality in terms of costs and benefits: quality increases based on the extent to which benefits outweigh costs. It is important to note that in this perspective the best-performing product or service may not deliver the best value and so may not be selected. Value and quality are often linked.

In translation, all five perspectives apply. A quality translation product must comply with the specifications agreed upon for production, such as the structured translation specifications introduced above, and those specifications must be appropriate to end-user needs. However, there is a tendency to emphasize transcendent quality, which has two main components, which are often called *accuracy* and *fluency*.[3] Accuracy refers to how well the source and target texts correspond. Fluency refers to how well the target text reads on its own as a document in the target language. A high-quality translation (from a transcendent perspective) would be perfectly accurate and fully fluent. However, there are two problems with the transcendent perspective in translation. First, there is often a tension between accuracy and fluency. For example, an overt translation may be quite accurate but lack fluency because it has a foreign feel to it, while a covert translation may be quite fluent yet not be accurate in an absolute sense because elements of the source culture have been converted to approximate equivalents in the target culture. Secondly, the transcendent perspective ignores specifications, other than the goal of perfect accuracy and fluency. In particular, a strictly transcendent perspective pretends that there is a perfect translation that does not

[3] Sometimes *adequacy* is used instead of *accuracy* but is more indirect and involves comparison with a reference translation. *Readability* is sometimes used instead of *fluency*, but is narrower in scope.

depend on audience, purpose, or cost. The value perspective can be used to avoid treating translation as a commodity, where cost per word is viewed as the only relevant factor (see Durban & Melby 2008).

1.6. DEFINING TRANSLATION QUALITY. In June 2012, I gave an invited workshop on translation quality at the NATO translation office in Brussels. The morning was spent discussing structured translation specifications and perspectives on quality. In the afternoon, after a lunch break, one highly experienced translator expressed her frustration by declaring that the transcendent perspective is the only possible perspective. She is not alone. There is a deep-seated and praiseworthy drive toward transcendent quality in all things. Perhaps your grandmother told you that if something is worth doing, it is worth doing well. Nevertheless, I suggest that the production and user perspectives should not be ignored. How can this dilemma be resolved? At the NATO workshop, the dilemma was resolved by distinguishing between *transcendent* quality and *functional* quality, where functional quality is focused on whether a translation works in a particular situation.

An element of resolving the tension between transcendent quality on the one hand and functional quality on the other hand is the recognition that just as no definition of translation is theory neutral, no definition of *translation quality* is theory neutral. There are many competing theories of translation. See Biguenet and Shulte (1992) for a compilation of primary sources illustrating translation theories from the 17th to the 20th centuries. See Gentzler (2001) for a presentation of more recent translation theories from the 1970s, 80s, and 90s. And see Pym (2010) for an analysis of both traditional and modern theories. As the primary basis for my definition of translation, I have chosen Functionalism, as promoted by Christiane Nord, and extended it by adding structured translation specifications as found in ISO/TS-11669 and ASTM F2575-14. Nord's Functionalism is not to be confused with Functionalism in philosophy, which sometimes takes a reductionist approach. Functionalism in translation is not reductionist. It acknowledges the variety of acceptable translations and types of translation alongside the vast array of unacceptable translations.

Functionalism in translation builds on the *Skopos theory* that developed in Germany during the 1980s. Also during the 1980s, unaware of Skopos theory, I developed my own theory of translation, based on a decade of work as a member of an interactive machine translation project team, followed by a decade of developing tools for human translators in the 1980s. I published an initial exposition on translation theory at the end of the 1980s (Melby 1990) but then discovered Functionalism and decided to work within it rather than compete with it.

Working within Functionalism means that we will assume that every translation has a purpose (that is, an intended function) and an intended audience, even if, in the extreme case, the audience is only the translator playing around with language and the purpose is only to have fun. Translation in a commercial environment is done for a requester with a purpose and an end-user profile in mind; otherwise, there would be no reason to pay for a translation.

Neither Functionalism nor any other translation theory is accepted by everyone. Although it seems obvious to talk about the purpose of a translation, relative to an intended

audience, it is a sufficiently significant shift from the past for it to be noted in the title of Nord's 1997 book, *Translation as a Purposeful Activity*.

As Ed Genztler (1998) points out in his review of Nord (1997), this is a shift from source-text oriented theories, where the only choice is between a "faithful" translation and a "free" translation of the source text. Nord "breaks the chain of 2,000 years of theory revolving around the 'faithful vs. free' axis" and brings in target-culture issues. For one audience and purpose a fairly literal translation might be more effective. For another audience and purpose, a free translation that replaces some foreign items of source culture with reasonably equivalent target-culture items might be more effective. The question is no longer whether all translations should be faithful to the original or whether all translations should be free translations. What used to be the fundamental question of translation theory becomes a strange question to even ask. The new question is what are the purpose and audience of the translation and how do they affect the translation "brief". I have replaced the *translation brief* with *structured translation specifications*.

1.7. A FUNCTIONALIST DEFINITION OF TRANSLATION QUALITY. At this point, I am prepared to present a definition of functional translation quality:

> A quality translation demonstrates the levels of accuracy and fluency required for the audience and purpose and complies with all other specifications negotiated between the requester and provider, taking into account both requester and end-user needs.

So far as I am aware, as simple as it looks, this is a new definition that has not been presented by anyone else. Its power may be hard to recognize because of its flexibility and simplicity.

This definition links to all five quality perspectives from the literature, except for value, which is implicit since price is addressed in the specifications. The definition as a whole is product oriented (talking about a translation and its attributes as well as the process followed to obtain it). Transcendent quality, while not accepted as an absolute, is acknowledged by explicitly mentioning accuracy and fluency.[4] Production quality is recognized as compliance with specifications, and user quality is brought in to remind us that specifications can be well-defined yet wrong for a particular audience.[5]

All three stakeholders are found explicitly in the definition. Requesters and providers negotiate the specifications, but those specifications must meet the needs of end-users. Structured specifications rescue the definition from undue vagueness.

Various types of assessment are supported by the definition, including:

---

[4] A translation so lacking in fluency that it is unintelligible or so lacking in accuracy that it fails to accomplish its purpose for the intended audience cannot be a quality translation, regardless of the specifications.

[5] Specifications can be wrong if they do not sufficiently describe the needs of the end user. Meeting wrong specifications obviously does not result in a quality product.

1. product assessment, in which case only the product-oriented translation parameters are considered,
2. process assessment, which focuses on the process-oriented translation parameters, and
3. project assessment, which focuses on the project-level translation parameters.

Suppose a highly-confidential German source text about a new type of light bulb is to be translated into Bulgarian and delivered to the client within three weeks, without divulging anything about it to the competition. Further suppose that the target text is translated according to the agreed-upon product specifications (i.e., parameters for source text and target requirements) but is delivered a month late and is useless to the client. According to a strictly product-focused assessment, the translation could be very good; but according to a project-focused assessment, it would be a bad translation outcome no matter how good the target text is linguistically, because the project violated an important relationship specification, the delivery date. Or suppose the translator asks a colleague for help and sends the confidential source text along with a question, but the colleague turns out to work for the competition, and the colleague passes on crucial information about the light bulb, and furthermore the competition takes advantage of the confidential information. According to a project-focused assessment, it would be a bad translation outcome because the confidentiality requirement of the environment specifications was violated.

Or suppose that the translation about the new kind of light bulb that was delivered to the client is eventually found to contain a serious error and the project manager wants to determine where the error was introduced. In this case, a process-focused assessment would be relevant. If records were kept of the target text after initial translation, after revision, after review, after final layout, and after proofreading, the project manager would be able to determine exactly where in the document production process the error was introduced and who missed it further along in the process, and corrective action could be taken.

The proposed definition supports any combination of product, process, and project assessment, depending on which categories of translation parameters are given values in the structured specifications.

In answer to the question of this section, translation quality turns out to be definable. Of course it remains to be seen whether this definition is useful. (Readers interested in further discussion of these definitions should consult Melby *et al,* 2014, Fields *et al*,2014, and Koby *et al*, 2014.)

2. ACHIEVABLE? Although much human translation is faulty, it is assumed in this paper that professional human translators can produce quality target texts, especially working into their native language on a text in a subject field they are familiar with. Thus, this section is not about whether high quality translation is possible in general but rather about the limits of *machine* translation and how to measure both human and machine translation quality. After mentioning some use-cases for machine translation and listing some highlights in the history of machine translation, a debate about its future will be summarized. Then the definition of translation quality developed in the previous section will be pre-

sented as a means to avoid endless back and forth about obstacles to machine translation and how they might be overcome.

It is difficult for many to accept that a translation does not always need to be completely fluent and accurate in order to be a quality translation. For example, a system that translates instant messages in real-time during a chat session between two people writing in different languages might include a machine translation system that exhibits functional quality even though it sometimes produces output that is unintelligible. Suppose that the specifications require that the system produce translations of up to 140 characters of source text within one second (this would be part of the value of ASTM F2575-14 Parameter 20, *Deadline*). If human translators take an average of ten seconds to translate an instant message (assuming any human translator would want to do this work), then they are not producing functional quality, even though their translations are substantially more accurate and fluent than raw machine translation. However, the specifications could also disqualify the speedy machine translation system if the requirement for accuracy (which falls under Parameter 9, *Content Correspondence*) and fluency (which falls under Parameters 6 and 12, *Target Language* and *Style*) is that fewer than one out of ten translations is rejected by the user as unintelligible, yet two out of ten chat translations are rejected by end users. An important empirical question is whether users are able to accomplish their objective of communicating via chat even though they don't both speak the same language. One factor that will be crucial is how often instant chat translations are accepted by an end user, that is, are reasonably fluent, yet cause serious misunderstandings because they are not accurate. If it turns out that interlocutors are able to detect fluent yet inaccurate translations and rephrase a question in order to solicit a response whose translation is more useful, then a fast machine translation system with less accuracy and fluency than a slower human translator would be seen as a quality solution while the human would not be seen as providing a quality solution. (While this way of discussing quality may be disconcerting to many readers, an alternative way of understanding its intent would be to see *quality* here as synonymous with an "excellent solution given the requirements" [Fields *et al.* 2014:411].)

Moving beyond this example of competition between humans and machines in a chat environment, consider environments where accuracy is paramount. There are a variety of translation specifications for projects involving health-care materials (where inaccuracies can result in harm or death to a patient) and transcription and translation of court depositions (where inaccuracies can result in a mistrial). Here speed does not compensate for lack of accuracy; and thus, these types of translation will have very different specifications than chat translation.

Consider what types of specifications would be appropriate for fully-automatic machine translation of technical support database entries, human translation of advertising materials, summary translations of messages between suspected terrorists, published translations of annual corporate reports, and even literary translation.

Some find it easier to accept a specification-based definition of quality, as opposed to a transcendent-only definition, by considering the fact that quality creates value. Three perspectives—*transcendent*, *production*, and *user*—work together to produce a product with

value to a requester or end-user. A product or service can have value without being perfect in all respects, but its value is related to how closely it meets agreed-on specifications.

People want things to become faster, cheaper, and better. Typically, vendors reply, "pick two of the three". We are told we must compromise, and we usually accept this. However, this dictum, applied to translation, assumes a transcendent view of quality, which is problematical, as explained above. In the proposed definition of translation quality, speed (meeting the delivery deadline) and cost are part of, rather than separate from, quality. Realistic specifications must balance speed, cost, and other aspects of quality. The question for translation then becomes whether a particular set of specifications is achievable and, if not, what compromises could be made to arrive at an achievable set of specifications. Many believe that productivity tools for human translators have allowed faster translation that is less expensive without loss of quality, but a careful study of this claim is beyond the scope of this paper, whose focus is machine translation.

A yearning for perfection has been behind the dream of machine translation from the beginning. Of course, for some sets of specifications, translation quality has already been achieved. For example, the European Commission and the United Nations each engage the services of many professional human translators to produce millions of words of quality translation per year, in that they satisfy their own specifications. However, these translations are expensive and slow. One aspect of the question for this section, "Achievable?" is whether it will ever be feasible to build a machine translation system that produces the ultimate in FAHQT: fully-automatic translation that meets requirements of accuracy and fluency tailored to the audience and purpose, and that is as good as the work of the best professional human translators, only much faster and much less expensive. In other words, can we have it all: faster, cheaper, and better (according to specifications that cannot be achieved by humans)?

In a sense, we have had it all with computers. For a number of years, computers have been getting faster, cheaper, and easier to work with. The key question is whether we can have a similar advance in machine translation. Clearly, such an advance would be a dream-come-true for machine translation software developers and a nightmare-come-true for human translators.

I would not dare offer an answer to this question; however, I will provide a little history of what others have said about it, and, in the conclusion I will comment on Kurzweil's prediction about what he calls the *Singularity* and how it might impact translation and other aspects of society.

2.1. HISTORY. The history of machine translation has been punctuated by periods of optimism and pessimism concerning this question. Peter Toma, one of the pioneers in commercial applications of machine translation, was involved in the early work in the 1950s and later founded Systran (http://www.systransoft.com/), a machine translation company that has been around for over forty years. Peter Toma was an optimist. He was not discouraged by the ALPAC report, and around 1975, in a face-to-face meeting, he told me that if someone would give him all the needed rules, he would program them and produce FAHQT. He was thus the epitome of the rule-based machine translation (RBMT) true believer. I was

hopeful that I would be able to help find some of those elusive rules that Peter was looking for. However, as we will see, rule-based approaches to machine translation have been largely overtaken by statistical approaches.

By 1980, I had gone through an intellectual crisis triggered by questioning my deepest beliefs about the nature of language. I switched from working on machine translation to working on productivity tools for human translators, with an emphasis on translator workstations, seeing the potential of microcomputers connected to remote computers as the home for translator tools, even before the introduction of the IBM PC in 1981. A detailed description of my intellectual crisis and revised views on the nature of language can be found in my book, *The Possibility of Language* (Melby 1995).

In 1980, while I was working on tools for human translators, Martin Kay wrote an internal report for Xerox PARC (Palo Alto Research Center) that was not published until 1997 but had a wide influence in the 1980s through photocopies that were passed around among those involved in translation technology. See Kay (1997[1980]), where he proposed a *machine-aided human translation* (MAHT) approach and argued against the use of raw machine translation.

Serge Perske led the machine translation effort in the European Commission. He held views diametrically opposed to those of Kay and me. In 1984, at Stanford University, during the Coling (computational linguistics) conference being held there, he approached me on a grassy campus quad and told me not to waste my time designing translator workstations, since within five years there would be no more human translators. A decade later, there were more human translators than ever, and Serge had lost his job. The massive rule-based machine translation project, Eurotra, was funded by the European Commission from 1978 to 1982. It never resulted in a commercially viable system that produced better results than Peter Toma's linguistically simpler approach.

In the early 1990s, a new approach to machine translation was initiated by a research project at IBM called CANDIDE. Based on the assumption that hand crafting thousands of rules to perform syntactic and semantic analysis of the source language, followed by adjustment of the resulting representation during the transfer phase, and then generation of a target-language text was a dead end, they tried a statistical approach. In the words of Adam Berger, a member of the CANDIDE team:

> In speaking a French sentence F, a French speaker originally thought up a sentence E in English, but somewhere in the noisy channel between his brain and mouth, the sentence E got "corrupted" to its French translation F. The task of an MT system is to discover [...] the optimal English sentence, given the observed French sentence. This approach involves constructing a model of likely English sentences, and a model of how English sentences translate to French sentences. Both these tasks are accomplished automatically with the help of a large amount of bilingual text. As wacky as this perspective might sound, it's no stranger than the view that an English sentence gets corrupted into an acoustic signal in passing from the person's brain to his mouth, and this perspective is now essentially universal in automatic speech recognition. (Berger 2013)

Although initially derided by the machine-translation community in the 1990s, within a decade, this approach, called Statistical Machine Translation (SMT), began to show some impressive results. Google Translate started out as a rule-based system, but in 2006 it was switched over to a statistical system (Gorman 2012).

Strange as it sounds, most of those involved in SMT development are not theoretical linguistics or translators. In fact, in most SMT systems, there is no syntactic analysis at all. Everything is based on simple sequences of words in source and target language texts. According to noted MT researcher Hans Uszkoreit (pers. comm.), it is not an uncommon phenomenon for SMT researchers to lose track of what languages they are working on while adjusting their systems.

Statistical machine translation systems are "trained" by feeding in a huge corpus of texts and their translations that have been pre-processed so that each segment of each source text is aligned with the segment of its target text that best corresponds to the translation of that source segment. From this corpus the system builds huge tables of correspondence showing how words will likely be translated when surrounded by particular words. I think of SMT as "massive parallel plagiarism" of bits and pieces from human translations and monolingual texts.

Also in 2006, the keynote presentation of a major machine translation conference was a public debate between a proponent of SMT, Daniel Marcu, and myself (a token linguist), about obstacles facing SMT, labelled as "data-driven machine translation" for the debate, and how to overcome them.[6] One of the slides from the debate features the following quote from a major figure in SMT:

> "Within the next few years there will be an explosion in translation technologies", says Alex Waibel, director of the International Centre for Advanced Communication Technology, in response to the question of how far machine translation systems can be taken. "There is no reason why they should not become as good, if not better, than humans", Dr Waibel says.

Daniel Marcu, my debate partner and a true believer in statistical machine translation, did not see the quote as farfetched.[7]

I suggested that several obstacles needed to be overcome by SMT in order for it to produce raw output equivalent to professional human translation. These obstacles included the need for machines to demonstrate the same competencies expected of a professional human translator, such as:

- Ability to understand the source text,
- Ability to write in the target language, and

[6] An annotated set of slides from that debate can be found at http://www.ttt.org/amta.

[7] In Waibel's 2014 LREC presentation accepting the prestigious Zampolli Prize he seems to soften this position.

- Ability to adjust to audience and purpose when translating and evaluating whether the source and target texts correspond

Daniel responded by noting that "airplanes don't bat their wings, but they still fly." In other words, there is no reason to believe that machines need to solve the problem of translation the same way humans do. He brought up the Chinese Room thought experiment[8] devised by John Searle, and suggested that for all we know, statistical machine translation systems may already understand what they are translating. As I have debated the question of machine understanding with other proponents of SMT, I have realized that this line of argument is fruitless. Until there is general agreement on what constitutes understanding and how to determine whether it is happening, there will be little progress in the debate as to whether machine translation systems understand language.

Another obstacle I presented was the need to take into account all types of context. In an article about context in translation (Melby & Foster 2010), I describe five types of context, and in the debate I pointed out that SMT then only dealt with a co-text (the words immediately surrounding a word) and bi-text (links between segments of source text and segments of target text). I suggested that machine translation systems will eventually have to deal with rel-text and non-text, that is, long-distance context where keyword lookup is insufficient, and context that goes beyond what is described in text. Daniel responded that given time and resources, machine translation will be able to reach beyond co-text and that there is no information needed by machine translation that cannot be found in texts. Thus, the debate about context led nowhere in 2006, except for an esoteric yet potentially significant question about "non-text" that can be examined by consulting the appendices to the Melby and Foster (2010) paper.

Yet another obstacle I presented is that in order to produce human-like translations, machines will need to demonstrate second-order creativity. Much has been written about orders of creativity. See, for example, Ekvall (1997). For the purposes of the debate, I defined first and second order creativity as follows:

> First-order creativity involves algorithmically generating an infinite number of items from a finite system; second-order creativity involves creating elements outside that infinite result.

Anyone familiar with Generative Grammar, as defined by Noam Chomsky and his many colleagues, will recognize that I am tying first-order creativity to what Chomsky has referred to as the *creative aspect* of language, namely, that a simple phrase-structure grammar that can be written down in a few lines can generate an infinite number of sentences using recursive relative clauses.[9] As applied to translation, an instance of second-order creativity would be either (a) finding a solution to a translation problem when that solution

---

[8] See http://plato.stanford.edu/entries/chinese-room/

[9] This sort of combinatorial process would be considered first-order creativity in translation, i.e., reusing pieces of existing human translations.

is not found in the corpus of bilingual texts that was used to train the system or (b) recognizing that none of the solutions in the corpus are appropriate. In other words, detecting that something is amiss in all the suggested translations and finding a solution not found in the material used for training the machine translation system would constitute second-order creativity. Daniel's response was that human translators make mistakes, so why should machines not be allowed to make mistakes?

The debate was inconclusive (an exit poll revealed that those in the audience who went into the debate believing in SMT as the ultimate solution still believed in it at the end of the debate and the skeptics remained skeptical). However, the rate of improvement in SMT output has slowed in recent years, and although I saw Daniel Marcu at a translation-related conference about five years after the initial debate and suggested another one, no date has yet been set for it.

As of 2016 (the last time anything but minor updates were made to this written version of a 2012 lecture), the verdict is still out on machine translation. There seems to be a growing consensus that statistical methods alone are not sufficient to achieve human-like translation, and a number of researchers are trying various hybrid techniques, combining statistical and rule-based approaches, or exploring the application of what are called "neural nets" to translation. Some high-profile commentators, such as Nicolas Ostler (2010) argue that it is inevitable that soon "everyone on the planet will be able to communicate using machine translation". The problem with statements like this is that they neglect to include a discussion of user requirements and the quality of machine translation relative to those requirements.

After careful consideration, I have concluded that it is not very useful to pile up obstacles to progress in machine translation. A more useful debate would be how to determine when a translation is a quality translation from a quality management perspective, i.e., according to an appropriate[10] metric. The definitions of translation and translation quality given in the previous section could be usefully debated. The definitions must be shown to be both reasonable and consistent with the real-world experience of at least some experts.

Once a group of people within the translation community (including machine translation developers, translators, translation project managers, translation requesters, end users, and academics) has reached consensus on definitions of translation and translation quality (similar to those presented in the previous section), it should become apparent that the question of whether quality machine translation is feasible is not a good question. It implies a yes or no answer. A much better question would be: how does one conduct a translation quality assessment?

The translation being assessed may be a raw machine translation, a human translation produced by one professional translator, or the result of a process that includes both machine translation and human translation, with several people involved and several types of technology in addition to machine translation.

[10] This formulation assumes the existence of multiple translation quality metrics, each tailored to particular requirements.

The first question to ask is "where are the specifications?" Assessment is a non-starter unless you have the specifications for the translation. If the specifications were not made explicit before the production phase began, then they must be inferred after the fact. Regardless of where the specifications come from, they must be appropriate to the needs of the expected end users.

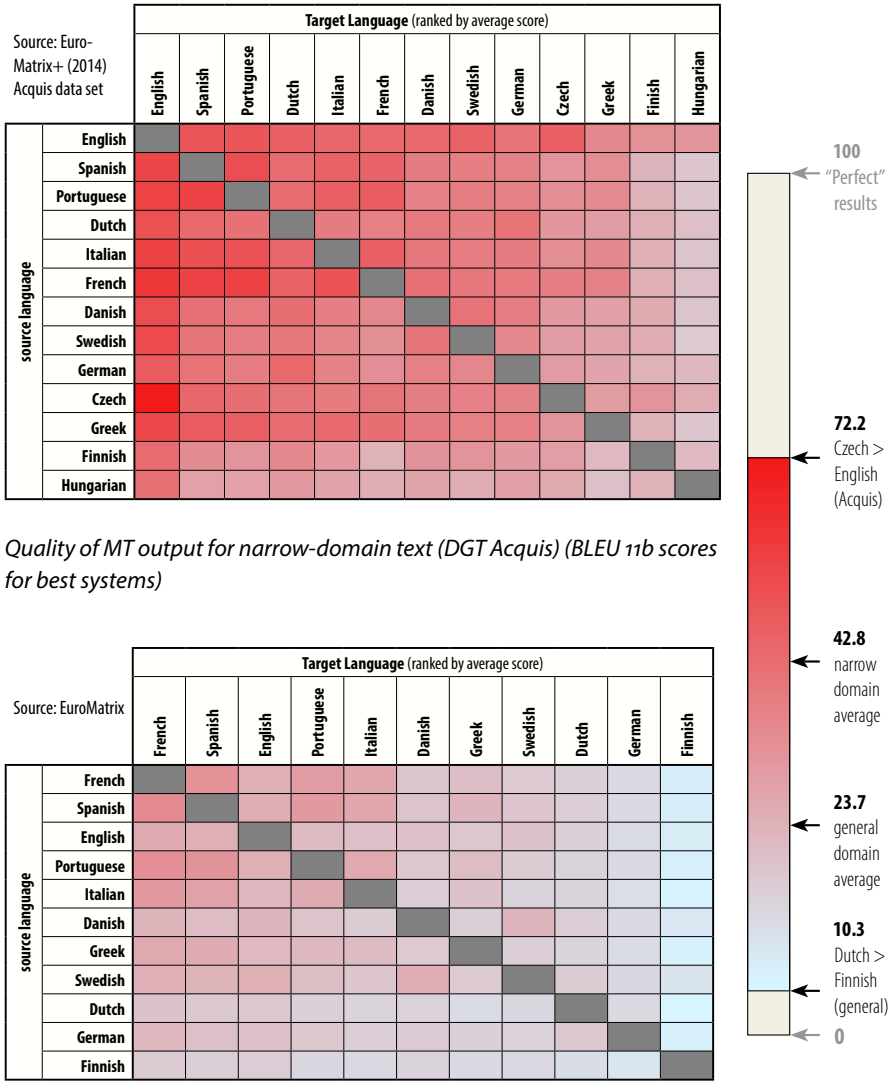
The second question to ask is "what is being assessed?" As previously discussed, assessment can focus on product, process, or project. Sometimes the product is assessed in isolation. Then, all the assessor has is the source text, the specifications, and the target text. Sometimes, aspects of the project will be assessed, such as whether specified resources were consulted, whether confidentiality was kept as required, and whether the product was delivered on time. And sometimes the process followed will be part of the assessment, especially when a product-oriented assessment reveals errors and the source of those errors is being tracked down. Project and process assessment are fairly straightforward, relative to product assessment.

When product assessment is involved, there are two well-known approaches: *holistic* and *analytic*.

A holistic assessment starts with the three essentials, that is, the source text, the target text, and the specifications, and the assessor grades the translation as a whole, relative to the specifications, along a scale such as good, bad, or mediocre. The problem with holistic assessment is that it is hard to know what to fix in order to improve. Actually, there is something even less useful than a holistic assessment as just described. That would be a holistic assessment without any explicit specifications. In that case, the grader uses whatever subjective criteria come to mind, without even knowing for sure why the translation was done or for what audience.

An analytic assessment starts with the same three essentials (source text, target text, and specifications) but uses a much more detailed metric, which typically includes error categories that are selected and weighted. Errors can be marked right down to the word level. The combination of errors and their severity produces a single number that represents how far the translation is from a quality translation; that is, one that meets the specifications, assuming the specifications are end-user appropriate. (Note that there are many possible metrics, each associated with a set of specifications. Thus, an approach to analytical assessment that is totally incompatible with the definition of translation quality in this paper is to attempt to define one metric that applies to all translations.)

Variations on typical analytic approaches to assessment are being explored, such as evaluation of pre-selected items (Kockaert & Segers, forthcoming), while various assessment methods used only for machine translation involve automatic comparison between raw machine translation and one or more "gold standard" translations of the source text, typically prepared by a panel of human translators, to see how similar the machine output is to the human translation. The machine translation system is tweaked and the output is compared with the gold standard. This is done over and over to see if modifications to the machine translation system improve or degrade the output. While such measures are useful, they are problematic because they depend on the reference translations. Preliminary research done at the German Research Center for Artificial Intelligence (DFKI), has



Quality of MT output for narrow-domain text (DGT Acquis) (BLEU 11b scores for best systems)

Quality of MT output for general-domain text (BLEU scores for best systems)

***Figure 1**. Machine translation quality of top-performing systems from the the EuroMatrix project, as measured by BLEU score. BLEU is a common MT quality evaluation metric that relies on the similarity of the translated text to human reference translations, where 100 is a "perfect" translation (i.e., one identical to a human reference translation), and 0 (zero) is a completely "bad" translation (i.e., it has* no *similarity to a reference). (Graphics taken from the QTLaunch-Pad project, created by Arle Lommel based on data from http://www.statmt.org/matrix/ and http://matrix.statmt.org, used with permission.)*

shown that when a translation's score depends on the set of references, changing the references used can change the score significantly without the translation under evaluation itself changing (Arle Lommel, pers. comm.). Nevertheless, these methods show that MT quality, if defined based on similarity to human reference translations, still falls short, even in the best of circumstances (see **Figure 1**, overleaf).

One project, called QT Launchpad (http://www.qt21.eu/launchpad/) has taken the definition of translation quality in this paper as a starting point and has developed a system for building customized metrics, called Multidimensional Quality Metrics (MQM) that are all tied to a product-oriented subset of the translation parameters discussed in this paper (see Lommel *et al.* 2014). It will be informative to see whether the metrics resulting from the QT Launchpad project can be shown to be valid and reliable in the sense these terms are used in assessment theory and practice.

A *reliable* metric is one that gives the same or very nearly the same result when used independently by two or more graders. In other words, it does not matter who grades a translation. The score will be about the same.

A *valid* metric is one that gives results that match the intuition of experts. One way to determine whether a translation metric is valid would be to give a team of translators who are recognized experts both the three essentials (source text, target text, and specifications) plus the score assigned to the translation by a grader. Validation could begin with some clear cut cases of good translations and bad translations. If a trained assessor using the metric assigns the good translations a good score and the bad translations a bad score, then the metric is at least somewhat valid. Less dramatically different translations could then be assessed to see if changes that the experts agree on are improvements, relative to the specifications, result in a better score, using the metric.

Clearly, a metric must be both valid and reliable in order to be useful. However, surprisingly, little has been done to study the reliability of translation quality metrics.

Now I can answer the question asked at the beginning of this section. Is quality machine translation achievable? Yes, sometimes. It depends on the structured translation specifications that are in force, the language combination, and the MT system.

For some specifications, quality machine translation was achieved long ago. One well-known example is the Météo system (Hutchins & Somers 1992:207–20) that was developed at the University of Montreal in the 1970s.

The specifications for the Météo system were to translate a particular type of Canadian weather bulletins from English to French well enough that only a word here or there needed to be corrected by a post-editor. All the bulletins were written by meteorologists who went through the same training, and they were all in the same style, resulting in what has been called a sublanguage. Météo was a spectacular success, in that most of the sentences of raw machine output were indistinguishable from what would have been produced by a professional translator, but it is hard to find additional sublanguages that are naturally occurring and have a continuous high volume of source texts that need to be translated. Nevertheless, Météo is an example of an environment in which quality machine translation has been achieved.

There are, of course, many examples of where, for a given machine translation system and a given set of specifications, including language combination, quality raw machine translation is not currently feasible. Combining humans and machines in HAMT or MAHT configurations brings many more possibilities.

Thus, the real question of this section is not whether quality translation involving machines is achievable but when it is achievable, always relative to explicit specifications.

3. DESIRABLE? The previous section attempted to demonstrate that quality translation involving machines is definitely achievable for some specifications. This is becoming less and less of an issue among those who adopt a functionalist view of translation quality, combined with the discipline of quality management. A question that is still an issue is whether quality machine translation is desirable. Should everyone focus on building a machine translation system that achieves FAHQT, that is, produces quality raw output no matter what the specifications? Although there is current government funding for machine translation development, it is for particular specifications, whether they are explicit or not. However, some academic researchers in machine translation are still aiming for FAHQT. Is this a waste of time and effort? For a number of years after my departure from machine translation development and my entry into translator tool development, I was uncomfortable with attempts to achieve FAHQT and actively discussed obstacles to it, implying that if researchers didn't know how to overcome those obstacles they should give up and pursue more promising research.

Recently, I have radically changed my position on the desirability of trying to achieve FAQHT. I now not only encourage work toward it but have devised a variation on the classic Turing Test that is intended to help us objectively determine whether it has been achieved in specific situations, and reduce the anxiety on the part of some human translators regarding machine translation as a potential threat to their livelihood.

In the classic Turing Test (Turing 1950), a computer attempts to imitate a human in a conversation.[11] The human judge is randomly connected to either a human or a computer for each test run. The two parties chat using instant messaging of some kind. In the 1950s this was accomplished using Teletype machines. At the end of each conversation, the human indicates whether the interlocutor was a human or a computer. If humans guess wrong more than 30 percent of the time, then the computer is deemed to possess human intelligence.

In my proposed variation, called the Translation Turing Test (TTT), a computer tries to imitate a professional translation project manager in a commercial environment. The human judge is a real customer, a requester of translation services. The requester develops an initial request for proposals (RFP) for a substantial translation project, using structured translation specifications as in this paper. Then the test proceeds like the classic Turing Test in that the customer is randomly connected to either a human or a computer. However, in the Translation Turing Test, the conversation is about the project specifications, rather than being a general discussion between strangers who may have little in common. The

[11] The title of the 2014 film *The Imitation Game* refers to Turing's own name for the test.

project manager (a human or computer) and the requester (a human) interact about the specifications in a back-and-forth conversation initiated by the requester,[12] clarifying and completing the specifications as needed, until the requester and the project manager have reached consensus concerning the specifications. Then, the conversation is put on hold until the agreed-upon due date, during which time the translation takes place (instant turnaround tasks, which obviously favor machine translation if the output is acceptable, are not included in this test). If the project manager is a human, s/he finds a qualified human translator who gets the job done using any desired technology, even translation memory and machine translation for selected segments. If the project manager is a computer, then absolutely no human involvement is allowed. The translation is then delivered to the customer, along with anything else listed in the specifications. Once again, the conversation is put on hold to give the customer a chance to contact a qualified translation assessment expert, who analyzes the translation according to the specifications. The customer then guesses whether the project was completed using humans or only raw MT. If the customer guesses wrong more than 30 percent of the time, the machine translation system is judged to have passed the Translation Turing Test.

Obviously, this is a very hard test to pass. Not only must the computer be capable of engaging in an intelligent conversation about translation specifications, which probably means it could pass the classic Turing Test (this is an interesting empirical research question), but the machine translation system must be capable of translating between any pair of languages in any subject field for any audience and purpose and satisfy all other elements of the specifications. Probably only academic researchers would attempt to pass the Translation Turing Test, yet it is clearly an interesting philosophical question whether a machine could be built that would do so.

What about those not interested in working on the Translation Turing Test? Are there less challenging but still useful goals to be pursued while the ultimate machine translation system is under development? Yes, definitely. Again the definition of translation quality in this paper helps define those goals.

I have devised another alternative to the Turing Test that brings together these goals. At a lunch meeting in Berlin with some colleagues, I became convinced that the second variation of the Turing Test strays too far from the classic test to be called a Turing Test, so I have named it the Translation Technology Test, to avoid any danger of posthumous discomfort for Alan Turing.

There are a multitude of Translation Technology Tests. Thus, we can distinguish from the Translation Turing Test (*the* TTT) and a Translation Technology Test (*a* TTT) by the choice of English article. Each Translation Technology Test is specific to one set of specifications that allow for a range of similar source texts and one translation environment, which may be anything from raw machine translation with no post-editing to human translation done by a human translator who uses a translation tool and whose competencies match the specifications. A translation quality metric is chosen that is appropriate for the specifications and was derived from a framework of metrics that has been shown to be *viable* (Snow, http://

[12] Conducted by instant messaging

scholarsarchive.byu.edu/etd/5593/). There is no conversation between the customer and the project manager during the test. A variety of source texts are submitted, and the translations are graded by translation quality assessment professionals. The test is not allowed to proceed unless that is a reasonable assurance that the grading is reliable. A threshold is established for grades according to how close professional human translators come to satisfying the specifications. If at least 70 percent of the translations meet the threshold (that is, if they fail less than 30 percent of time, without considering turnaround time), then that translation environment is judged as having passed one Translation Technology Test, the one associated with those specifications. If the turnaround time requirements are such that human translators cannot meet them, the question reduces to whether raw MT can meet the specifications. It must be kept in mind that the threshold of meeting the specifications is not set according to whether the translation meets transcendent requirements of full accuracy and fluency. The threshold is set according to whether the translation meets the specifications, which may, for example, de-emphasize fluency. This implies that the metric must be built according to the specifications, which is consistent with the principle that there can be no universal translation metric.

TTTs apply to the entire spectrum of type of translation described at the beginning of this paper (from fully automatic, to HAMT, to MAHT[13]). They also apply to human translation where no technology is used other than a pencil and paper, although this type of translation is no longer competitive with MAHT in the commercial market. Its only place is in a translator certification examination, and even there, some translator certification systems are moving toward keyboarded examinations on a computer.

In some ways, Translation Technology Tests have been around for a long time. As soon as people started backing off from seeking one metric to assess all kinds of translation environments and began believing that machine translation can be useful even if it is not fully accurate and fluent, they essentially started moving toward Translation Technology Tests (TTTs). What I am suggesting is that a formal and repeatable methodology for TTTs should be developed and used extensively throughout the translation industry, including government translation offices, and even in academia for translator education.

In answer to the question of this section, it would be desirable to develop a family of valid and reliable TTTs (Translation Technology Tests) that determine whether particular combinations of human and machine translation meet real-life translation needs.

4. CONCLUSION. As described in section 2 ("Achievable?"), predictions of the imminent replacement of all human translators by machines have come and gone, often with a promise of significant progress within five years. Ray Kurzweil, a futurist, has been writing about the upcoming *Singularity* for years. The Singularity is the point in time when human intelligence is given to a computer, probably an embodied computer. Kurzweil's 2005 book *The Singularity is Near*, predicts that the Singularity will take place by the year 2029. In a June, 2011, interview with Ray Kurzweil, Nataly Kelly (2011) confirmed with Ray Kurzweil that

[13] For example, such a TTT could be used to compare student and professional translators.

the 2029 Singularity includes the capability to translate as well as or better than the best human translators. In other words, we will have FAHQT and a machine will pass the Translation Turing Test by 2029 (or perhaps by 2045, according to http://www.singularity.com/), according to Kurzweil

I have read *The Singularity is Near*, and I have come away from it convinced that at least Kurzweil has a feeling for how difficult the task of replicating human intelligence will be. Nevertheless, I do find what I think is a major flaw in Kurzweil's logic. He predicts that once we have built true artificially intelligent entities, they will be able to share knowledge much more quickly than humans can, essentially instantly. I strongly disagree. Kurzweil acknowledges that Singularity intelligences would be able to reconfigure themselves and develop new conceptual categories at the most basic level of fundamental world view, just like humans, at least open-minded humans. Instant information sharing within a predetermined, shared domain of knowledge may be possible, but each Singularity intelligence will quickly evolve as it learns and integrates perceptions into unique mental categories, just as each human intelligence is unique. Singularity intelligences, just like people embedded in cultures, would each evolve along their own path and would probably have to use language, with all its inefficiency and fundamental ambiguity, to communicate among themselves. Thus, information sharing between Singularity Intelligences will be slow and error prone, just as it is between humans.

As pointed out by a colleague, Robert Orr, after the 2012 oral version of this paper, what sets humans (and, I would add, Singularity intelligences) apart from current statistical machine translation systems is their ability to organize a collection of instances of language, such as idioms, into a coherent semantic system without a pre-existing framework. Orr cites Makkai (2009),which proposes a two-stage development of idioms in child language acquisition, as an example of the distinction between what statistical machine translation can currently do and what humans can do.

Ultimately, I believe, machine translation systems will need to organize language information into the same strata used by the human mind. See Lamb (1999) for a detailed exposition of these strata from multiple perspectives (linguistic, cognitive, and neurological).

In addition to the ability to organize information the way humans do, Singularity intelligences would necessarily possess free will, as do humans, and might choose other activities than translation. Thus, it is an open question whether Singularity intelligences would choose to replace human translators. Kurzweil himself suggests that the human translation profession would evolve but not disappear.

Since Kurzweil has been bold enough to make a prediction about the future of machine translation, I will also make some predictions, based on my deeply-held conviction, stated in the introduction to this paper and shared by many others, that professional translation is one of those intellectually challenging activities that set humans apart from all machines other than truly intelligent machines (which may not be around until 2046 or beyond). My three predictions are:

1. The acceptance of my definition of translation quality and the use of structured specifications in particular will gradually increase and will help all stakeholders

communicate better, and both assess and improve human and machine translation quality through human-machine collaboration, not competition.

2. During the rest of the decade, that is, between 2012 and 2020, the only humans who will be replaced by MT will be those who translate mechanically, without fully understanding the source text or without substantial domain knowledge and target-language writing skills. Currently, raw machine translation is an alternative to "zero translation" (no translation at all), rather than an alternative to professional human translation.
3. If Kurzweil is right and the Singularity arrives by 2045, the news that machines can translate as well as professional human translators will be lost in the world-wide commotion brought about by the realization that computers have become smarter than humans. This will impact every aspect of human life, since every interaction by text or voice could be with a human or a machine. Almost everyone on the planet will be in danger of losing their job to a robot, assuming the raw materials and energy needed for them to build more of themselves will be available. I personally look forward to this adventure and hope to be around to see what happens when the Singularity arrives, hopefully sooner rather than later, since I will be nearly a hundred years old in 2045.

The bottom line is that human translators should stop worrying about being replaced by a computer and instead make friends with translation technology. Buyers of translation, likewise, should engage the services of professional human translators, except in particular circumstances where raw machine translation can be shown to pass a Translation Technology Test based on applicable specifications. Without the Singularity as predicted by Kurzweil, I claim there will always be plenty of interesting work for high-end professional translators, since only humans will be able to deal effectively with the most interesting and challenging sets of specifications. Now, in practical conclusion, I present serious action items for five types of players in the translation industry (from my 2012 lecture):

1. for *requesters and providers*: use structured translation specifications all along the way from provider selection to production to assessment of the results;
2. for *translation tool developers*: continue to focus on human-machine collaboration and incorporate translation quality metrics into tools;
3. for *machine translation developers*: embrace the idea of Translation Technology Tests as described in this paper;
4. for *AI researchers*: aim at passing the general Translation Turing Test (but beware of autistic MT!—some explanation of this comment is required for further discussion, but that is for another paper); and
5. for *end users*: ask for the specifications that were used during translation and make use of them in assessment and in providing feedback.

REFERENCES

ALPAC. 1966. *Languages and machines: computers in translation and linguistics. A report by the Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Academy of Sciences, National Research Council*. Washington DC: National Academy.

BELLOS, DAVID. 2011. *Is that a fish in your ear?: Translation and the meaning of everything*. London: Faber & Faber.

BERGER, ADAM. 2013. Statistical Machine Translation. http://www.cs.cmu.edu/~aberger/mt.html (accessed August 2013).

BIGUENET, JOHN & RAINER SCHULTE (eds.). 1992. *Theories of translation: An anthology of essays from Dryden to Derrida*. Chicago: University of Chicago Press.

DURBAN, CHRIS & ALAN K. MELBY. 2008. *Translation: Buying a non-commodity* (brochure). Alexandria VA: American Translators Association.

EUROPEAN COMMISSION. 2012. *Quantifying quality costs and the cost of poor quality in translation*. EC document ID: HC-31-12-463-EN-C. http://bookshop.europa.eu/en/quantifying-quality-costs-and-the-cost-of-poor-quality-in-translation-pbHC3112463/ (accessed August 2013).

EKVALL, GÖRAN. 1997. Organizational conditions and levels of creativity. *Creativity and innovation management* 6(4):193–258. (DOI: 10.1111/1467-8691.00070).

FIELDS, PAUL, DARYL HAGUE, GEOFFREY S. KOBY, ARLE LOMMEL & ALAN MELBY. 2014. What is quality? A management discipline and the translation industry get acquainted. *Tradùmatica* 12:404–412.

GENTZLER, EDWIN. 2001. *Contemporary translation theories*. Bristol: Multilingual Matters.

———. 1998. Review of *Translation as a purposeful activity: Functionalist approaches explained*, by Christiane Nord. Translation and literature 7(2):266–76.

GLEICK, PETER H. 2000. *Harry Potter, minus a certain flavour*. Op-ed, *New York Times*, July 23.

GOLDSTEIN, STEVEN. 2004. Translating Harry, part I: The language of magic. http://bytelevel.com/global/translating_harry_potter.html (accessed June 2014).

GORMAN, MICHAEL. 2012. Google gives us some insight on the inner workings of Google Translate. http://www.engadget.com/2012/03/19/google-translate-how-it-works/ (accessed July 2013).

HAGUE, DARYL, ALAN MELBY & WANG ZHENG. 2011. Surveying translation quality assessment: A specification approach. *The interpreter and translator trainer* 5(2):243–67.

HARRIS, ROY. 1982. *The language myth*. Longon: Duckworth.

HOUSE, JULIANE. 1997. *Translation quality assessment*. Tübingen: Gunter Narr.

———. 2001. Translation quality assessment: Linguistic description versus social evaluation. *Meta: Translators' journal* 46(2):243–57. http://id.erudit.org/iderudit/003141ar (accessed August 2013).

———. 2010. Overt and covert translation. In *Handbook of translation studies*, vol. 1, ed. by Yves Gambier & Luc van Doorslaer, 245–46. Amsterdam: Benjamins.

Hutchins, John & Harold L. Somers. 1992. *An introduction to machine translation*. London: Academic Press. (Relevant chapter is available online at http://www.hutchinsweb.me.uk/IntroMT-12.pdf).

iso. 2012. *iso/ts-11669:2012, Translation Projects -- General Guidance*. Geneva, Switzerland: International Organization for Standardization.Jakobson, Roman. 2000[1959]. On linguistic aspects of translation. In *The translation studies reader*, ed. by Lawrence Venuti, 113–18. London: Routledge.

Kay, Martin. 1997[1980]. The proper place of men and machines in translation. *Machine translation* 12:3–23.

Kelly, Nataly. 2011. Interview with Ray Kurzweil. http://www.huffingtonpost.com/nataly-kelly/ray-kurzweil-on-translati_b_875745.html (accessed August 2013).

Koby, Geoffrey S., Paul Fields, Daryl Hague, Arle Lommel & Alan Melby. 2014. Defining translation quality. *Tradumatica* 12:413–420.

Koehn, Philipp. 2010. *Statistical machine translation*. Cambridge: Cambridge University Press.

Kockaert, Hendrik J. & Winibert Segers. 2014. Evaluation de la traduction : la méthode PIE (Preselected Items Evaluation). *Turjuman: revue de traduction et d'interprétation / journal of translation studies* 23 (2): 232-250.

Kurzweil, Raymond. 2006. *The Singularity is near: When humans transcend biology*. New York: Viking.

Lamb, Sydney. 1999. Pathways of the Brain: The neurocognitive basis of language. Amsterdam: John Benjamins.

Lommel, Arle, Aljoscha Burchardt, Maja Popović, Kim Harris, Eleftherios Avramidis & Hans Uszkoreit. 2014. Using a New Analytic Measure for the Annotation and Analysis of MT Errors on Real Data. *Proceedings of EAMT 2014*.

Makkai, Adam. 2009. On Redefining the Idiom *LACUS, forum* 36:215–27.

Melby, Alan. 1990. The mentions of equivalence in translation. *Meta: Translators' journal* 35(1):207–13.

———. 1995. *The possibility of language*. Amsterdam: John Benjamins.Melby, Alan & Christopher Foster. 2010. Context in translation: Definition, access and teamwork. *Translation & interpreting* 2(2):1–15.

Melby, Alan, Alan Manning, & Leticia Klemetz. 2007. Quality in translation: A lesson for the study of meaning. *Linguistics and the human sciences* 1(3):403–46.

Melby, Alan, Paul Fields, Daryl Hague, Geoffrey S. Koby & Arle Lommel. 2014. Defining the landscape of translation. *Tradùmatica* 12:392–403.

Murphy, Cait. 1995. "Ulysses" in Chinese: The story of an elderly pair of translators and their unusual bestseller. *The Atlantic*, September. http://www.theatlantic.com/past/docs/issues/95sep/ulyss.htm (accessed June 2014).

Nord, Christiane. 1997. *Translation as a purposeful activity*. Manchester: St Jerome Publishing.

Ostler, Nicholas. 2010. *The last lingua franca*. New York: Walker & Company.
O'Sullivan, Carol. 2013. Introduction: Multimodality as challenge and resource for translation. *Jostrans: The journal of specialized translation* 20:2–14.
Parameters. 2013. Structured translation parameters, version 6. http://www.ttt.org/specs (accessed August 2013).
Pym, Anthony. 2010. *Exploring translation theories*. London: Routledge.
Radosh, Danie. 1999. Why American kids don't consider Harry Potter an insufferable prig. http://www.radosh.net/writing/potter.html (accessed June 2014, originally published in *The New Yorker*, Sept. 20, 1999).
Snow, Tyler. forthcoming. Establishing the viability of the Multidimensional Quality Metrics framework. MA thesis, Brigham Young University.
Tolkien, J.R.R. 1965[1955]. *The return of the king*. New York: Ballantine.
Turing, Alan. 1950. *Computing machinery and intelligence*. Mind 59(236):433–60.
Venuti, Lawrence. 1998. *The scandals of translation: Towards an ethics of difference*. London: Routledge
———. 2009. Translation, empiricism, ethics. Paper presented at the 2009 MLA Conference, Philadelphia, December 28.
Weaver, Warren. 1955[1949]. Translation. In *Machine translation of languages: Fourteen essays*, ed. by William N. Locke & A. Donald Booth, 15–23. Cambridge MA: Technology Press of the Massachusetts Institute of Technology.