This presentation discusses the European Union-funded QTLaunchPad project, and particularly the Multidimensional Quality Metrics, or MQM, framework for translation quality assessment.
QTLaunchPad is an EU-funded project that focuses on the barriers to high-quality translation. While much of the focus is on MT, one of our core principles is that language technology should utilize the strengths of humans and machines and not try to cut humans out of the process. Achieving this goal is a major undertaking, which we have begun with QTLaunchPad.
Many ways of assessing quality = no agreement

There are currently many ways of assessing translation quality, ranging from very informal to highly formal. The results of these methods are often not comparable, which poses a problem when trying to compare quality results from different translation methods.
Research methods such as BLEU and METEOR that rely on reference translations are useful for development tasks where the ability to quickly determine how changes to systems work is important. But critically, the results may actually not mean anything for human users. In the example shown above a system change results in a BLEU score improvement of 3 points by improving the low (the “bad” to “really ugly”) end of the quality spectrum.
While in this case there is no BLEU improvement even though the “good” and “almost good” portion of the translated text that is useful to humans has increased substantially and the distinction between the good and bad content has become clearer.
MT methods require reference translations:
Cannot be used for production purposes

More importantly, reference-based measures cannot be used in a production environment because they require that the text have been previously translated, a condition that cannot be met for newly translated text.
On the other hand, human assessment takes too long and far too often is based on the preferences of individuals. Anyone who has been involved in quality assessment sooner or later comes across a case where two reviewers disagree strongly about whether a text meets quality expectations.
Furthermore, we often do not know what we mean by quality and disagree as a result. If we cannot agree on what we mean by quality, how do we know when we have achieved it? Beyond that, how can we account for the fact that translations serve different needs, and the standard for a corporate annual report will be very different from that used for an in-house automotive service manual?
In response to this problem, we propose a functionalist definition of quality, originally proposed by Alan Melby and refined in the course of the QTLaunchPad project:

A quality translation demonstrates required accuracy and fluency for the audience and purpose and complies with all other negotiated specifications, taking into account end-user needs.

While the details are beyond the scope of this presentation, what is important to note is that this definition maintains that there actually is no single set of quality criteria that apply in all cases: instead quality must be sensitive to purpose and situation.
And that definition, of course, addresses why we can’t have a single metric. But even if we could, we quickly face the question…
…of which metric we should use. There are many metrics out there, ranging from simple to complex, with many different purposes. On this slide you can see just a few of them. We examined all of these and others and…
…were surprised to find that there was only a single issue that all of them agreed upon: correct terminology. Although each of them is put forward as a measure of “quality”, each of them actually measures something different. But based on our definition of quality, we can understand why: they serve different needs.
So what is the solution?
In our project we created a master listing of all of the types of issues checked in different metrics. The result is a huge hierarchy, which you can see here. We like to call this the “bowl of spaghetti.” So the correct response to seeing this is...
Are you kidding?! We realize this is too much for anyone to use. But each of the items you saw is actually used by someone.
Realizing that is too complex, we have created a simpler “core” with 21 issues that are more commonly addressed. We don’t expect anyone to use all of even this, but by limiting one’s choices to the core, you increase the likelihood of interoperability.
Don’t assume. Use specifications.

But how do you know what to check from that huge list? The important and innovative aspect of MQM is that, rather than assuming what to check, you decide what to check based on specifications about the translation that are agreed upon by the requester and provider.
These specifications are based on 12 parameters adapted from the ASTM F2575 and ISO/TS-11669 standards. We will go through them in more details tomorrow in the quality assessment workshop. By defining what your expectations are in advance, even before you request a translation, you can ensure that the quality is evaluated in response to your needs.

<table>
<thead>
<tr>
<th>Specifications based on 12 Parameters</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Specifications based on 12 Parameters</td>
<td>(from ISO/TS-11669)</td>
</tr>
<tr>
<td>1. Language/locale</td>
<td>7. Register</td>
</tr>
<tr>
<td>2. Subject field/domain</td>
<td>8. Target text style</td>
</tr>
<tr>
<td>3. Terminology (source/target)</td>
<td>9. Content correspondence</td>
</tr>
<tr>
<td>4. Text type</td>
<td>10. Output modality</td>
</tr>
<tr>
<td>5. Audience</td>
<td>11. File format</td>
</tr>
<tr>
<td>6. Purpose</td>
<td>12. Production technology</td>
</tr>
</tbody>
</table>

http://www.ttt.org/specs
With MQM, therefore, you use only those parts of the hierarchy that check what your specifications define as important. For example, if style doesn't matter for your audience, you don't check it, but if it does, then you can include it.
MQM lets you declare your quality metric in a shared vocabulary.

By using a shared vocabulary, you make sure that everyone knows what is to be checked and you define your terms for assessing quality up front.
For example, here is metric we developed specifically for assessing the performance of MT systems. (Note that it includes a few extra bits under “Function words”: MQM is extensible when you need it to be, although obviously, for interoperability, you should use established issue types when possible.)
And here is an MQM representation of the SAE J2450 standard automotive service manuals. As you can see, this looks very different from the MT-centric metric, but defining them in terms of MQM allows us to see the differences and better understand them.
We have developed a number of tools for use with MQM, ones that will help use to define and try different metrics within the MQM framework. These are available from the URLs shown.

DEMO: http://www.translate5.net
http://scorecard2.gevterm.net
So, what advantages does MQM offer for large organizations? There are a number:

1. Many of the MQM categories are potentially automatable using language technology. By selecting automatic issues, you can arrive at metrics that can be used to give you an automatic score.
2. Organizations can agree upon shared MQM metric profiles for different tasks, which can help them benchmark and compare quality and work towards collective improvement.
3. By using automatic issues, you can also improve sampling methods so that they focus on portions of the text that are likely to cause problems.
4. By switching to metrics defined in MQM you can make MT and human translation comparable and actually see where each one works well. Without this comparability the business decision about which to use is difficult.
5. You can include your quality profile in RFPs, which lets LSPs know in advance what they will be expected to provide, resulting in more accurate bids and smoother relationships with translation providers.

Of course the benefits are not limited just to large organizations, but they are often the ones experiencing the most critical need for quality assessment.
At this stage we are looking for feedback and for individuals interested in testing or piloting MQM. We are already working with translation tool providers to see MQM added to their tools and it has been piloted with a number of larger organizations, but we need more feedback. The workshop tomorrow provides you with an opportunity to learn more and see MQM in action. The URL on screen is where you can go for the formal definition of MQM.

Looking for feedback:

http://www.qt21.eu/mqm-definition