Unbabel

AMTA 2022

Business Critical Errors: A Framework for Adaptive Quality Feedback

Craig Stewart
Research Scientist, Unbabel

Marianna Buchicchio
Senior NLP Quality Analyst, Unbabel

Madalena Gonçalves
Junior NLP Quality Analyst, Unbabel

Alon Lavie
VP Language Technologies, Unbabel

Table of contents

01

Motivation

02

What is BCE?

03

Applications

Introductions

Craig Stewart

Research Scientist

Marianna Buchicchio

Senior Quality Analyst

Madalena Gonçalves

Junior Quality Analyst

Alon Lavie

VP of Language Technologies

The Unbabel team

Roles & Responsibilities

The traditional landscape in translation

Machine-only

Lacks the necessary quality
for a reliable customer experience

VS

Human-only

Does not scale
to the growing mountains of digital content

Unbabel’s Translation Platform

AI Stack

Community

Proprietary data

Continuous learning

Seamless Integrations

Motivation

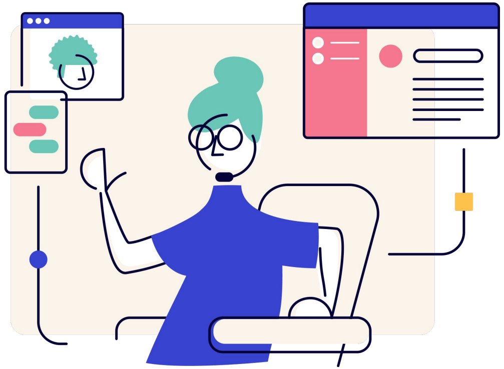
What is a ‘good’ translation?

In many cases, customer expectation can deviate from linguistic quality. Nuanced brand requirements, for example, can render perfectly sound translation ineffective for a specific use case:

What if a customer wants all of their content written in lower case?

What if they want to mix formal pronouns with a more informal discourse style?

Quality expectations can be both objective and subjective



Confidentiality level: External Use

What is a ‘good’ translation?

For this reason, at Unbabel we approach quality on **two dimensions**:

Linguistic Quality

To what extent is the translation linguistically accurate?

For us, at Unbabel, Multidimensional Quality Metrics* (MQM) is the most useful measure of linguistic accuracy.

We adapt the framework to align with our use cases.

*http://www.qt21.eu/mqm-definition/definition-2015-12-30.html

Utility

To what extent is the translation ‘fit for purpose’?

MQM can capture some of this information and there are strategies for adapting MQM to customized requirements such as weighting systems on top of severity multipliers.

There is a growing need for leveraging MQM in different ways to accommodate variable expectations.

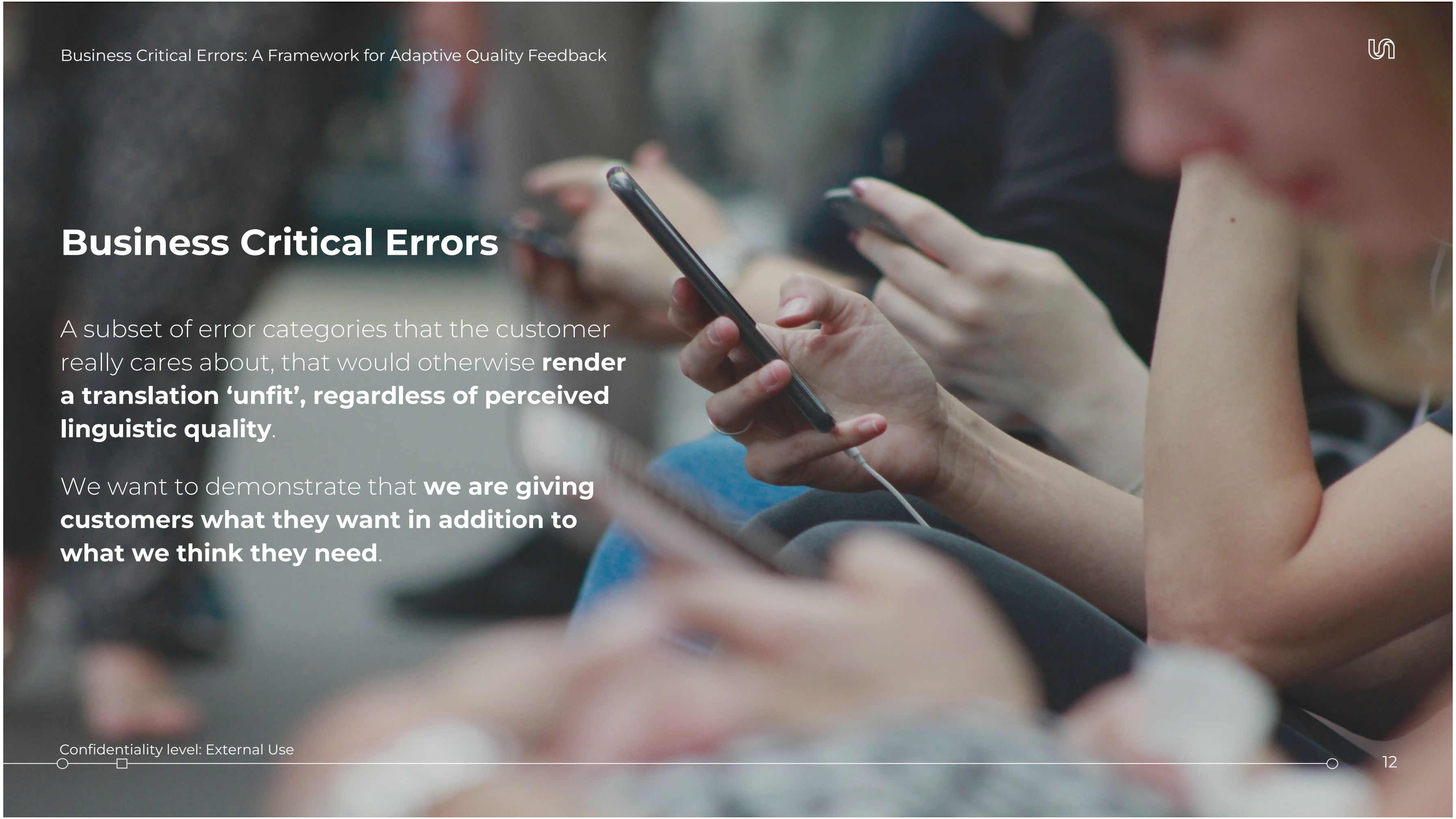
Unbabel is built on quality agility

We service the widest possible range of quality expectations from synchronous customer chat to on-brand marketing content.

We need a quality evaluation solution which can accommodate all expectations

MQM has been pivotal in allowing us to leverage an in-house community combined with a suite of AI evaluation tools which enable us to be highly adaptive. But we believe we can go further...

What is BCE?



Business Critical Errors

A subset of error categories that the customer really cares about, that would otherwise **render a translation ‘unfit’, regardless of perceived linguistic quality**.

We want to demonstrate that **we are giving customers what they want in addition to what we think they need**.

Business Critical Errors

Objectives

Expressivity

Articulating adherence:

We want the framework to adequately express how we are meeting expectations (or not!)

Efficiency

Minimize extra overhead:

Ideally we don't want to have to add any extra work for annotators or complicate and slow down the evaluation process

Simplicity

Minimize complexity:

Adding extra dimensions to MQM can make it difficult to interpret consistently.

Business Critical Errors

Approach

Expressivity

Figure out which error types the customer really cares about

Define priority error types that can be broadly applied and are impactful

Efficiency

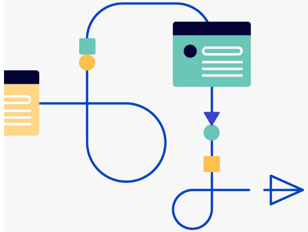
Use the existing framework and ring fence a subset of errors

We only have to make a single pass of annotation with minimal special instructions to the annotator.

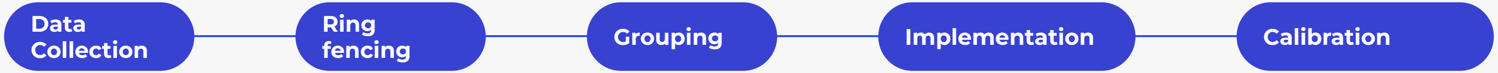
Simplicity

Define a minimalist set of error types

Report on counts of occurrences of BCE type errors and isolate that calculation from MQM.



Defining the framework



Data Collection

Gather feedback from customers, both from interview and existing complaints

Ring fencing

Use distribution of collected data to establish the most critical error types

Grouping

Establish a minimal set for groups of content relative to quality expectations

Implementation

Develop tooling for pulling counts of BCE from annotations and for reporting

Calibration

Working with customers to refine the categories, monitoring business impact

BCE as a Metric

How do we turn counts of these errors into a measurable metric?

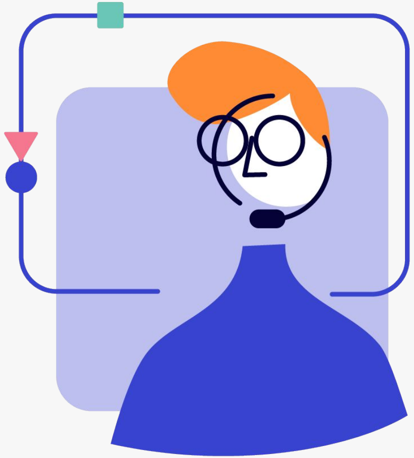
We currently define our BCE metric as **the number of BCE errors per 1000 words**.

This is implemented such that **we can generate the metric once per quarter** in order to track progress over time and demonstrate improvement.

Why not just weight MQM scores?

We want this to be adaptive, so having different MQM values per customer would cause confusion

How has this been useful to us?



Allows us to prioritize

The biggest benefit is in **tightening our feedback loops** and allowing us to **focus on the issues that really matter**. Rather than sifting through all of the issues we can discover the issues that will have the greatest impact on the customer.

Quality Agility

With minimal overhead, we are now **able to customize quality feedback in meaningful ways** and show the customer that we really know and understand their expectations.

Improved processes and tooling

BCE generates an extra source of data that can complement our internal processes and tooling. We can **evaluate our MT models** specifically on BCE and **develop Quality Estimation models** focused on high impact error.

Applications

Applications of BCE at Unbabel

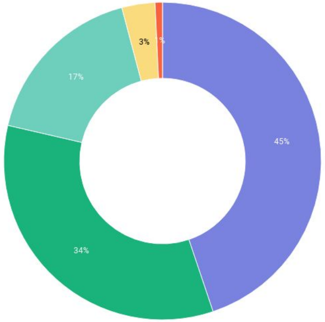
As we refine the framework we have found specific use cases in which we can use it to improve our tooling and processes:

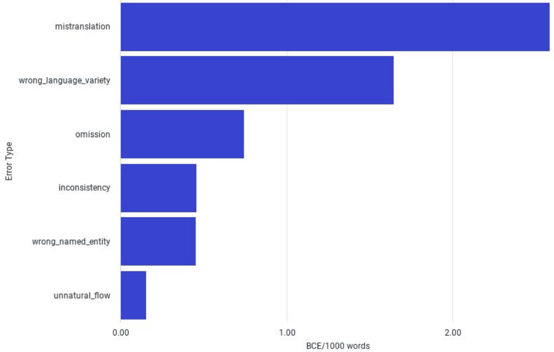
Customer Utility Analysis

The primary intention for BCE is to **complement customer reporting**.

Our **Customer Utility Analysis Framework** allows us to clearly communicate the quality of translation.

We report **linguistic quality relative to distributions of bucketed MQM scores** which can be accompanied by our **BCE metric for translation utility**.





Confidentiality level: External Use

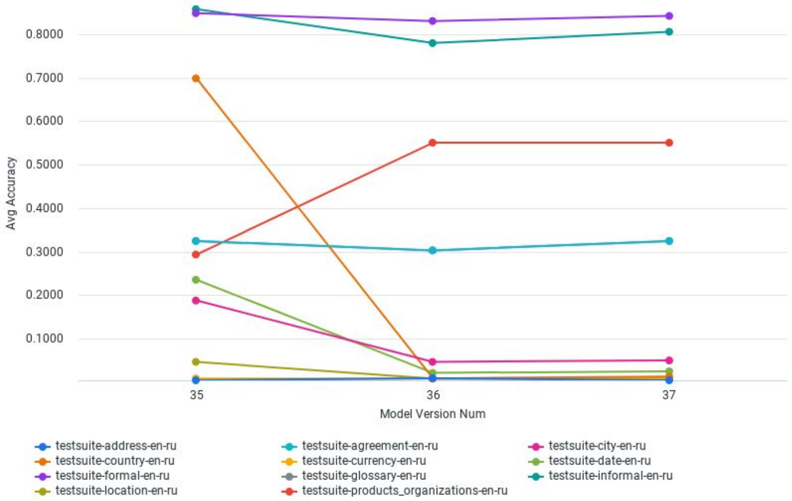
Applications of BCE at Unbabel

MT Model Evaluation

We have developed **BCE Test Suites**; benchmarking test sets by which we **evaluate the performance of our MT systems on specific phenomena**.

We put our **MT models through a gauntlet of specialized test sets** by which we established their ability to avoid certain BCE.

In this way we can **maximize translation quality downstream in meaningful ways**.



Applications of BCE at Unbabel

Automated Metric Evaluation

Our homegrown automated evaluation **metrics (COMET) are also tested for their ability to capture BCE**.

Similarly to MT systems, we have developed a gauntlet of test sets whereby **we ask our metrics to rank segments to ensure that the segment containing BCE receives a lower ranking**.

Applications of BCE at Unbabel

Quality Estimation

We have developed **specialized Quality Estimation systems that are trained on BCE data** and **predict the number of BCE errors per segment**.

We can use these systems as **a flagging mechanism to catch BCE before it goes out the door** and reroute it for human review.

OpenKiwi

By **Unbabel**

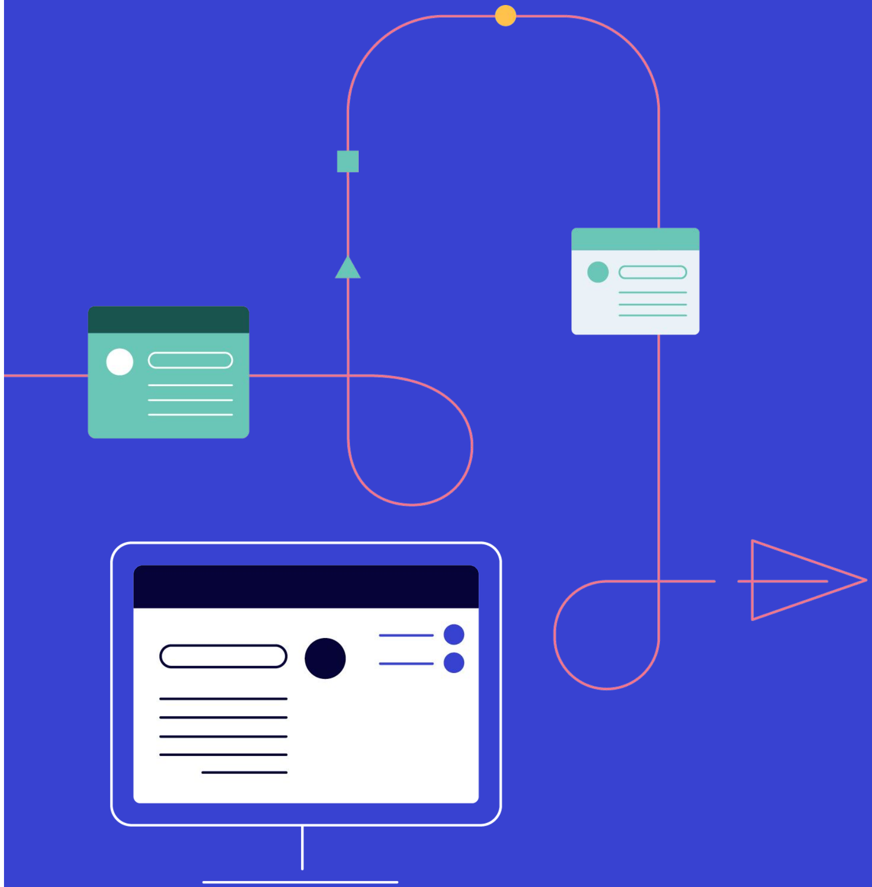
Summary

Key Takeaways

Quality expectations can be both **objective and subjective**

Business Critical Error (BCE)…

- is **focused on subjective expectation**
- allows us to **give customers what they want** vs **what we think they need**
- enables us to **prioritize issue resolution**
- can help us **design translation solutions that fit particular dimensions**
- provides **a rich source of high-impact data**



Questions?

Unbabel