GILT Metrics – Slaying the Word Count Dragon

In the beginning...

One of the most enduring features of the GILT (Globalization, Internationalization, Localization and Translation) industry has been the inconsistency of word counts, not only between rival products, but even between different versions of the same product. Trying to establish a measure for the size of a given GILT task is not unlike trying to fight a many-headed dragon.

The havoc that the lack of a uniform system of measurement can cause was recently exemplified in 1999 when the Mars Climate Orbiter Spacecraft was lost because one NASA team used Imperial units, while another used metric units, for a key spacecraft operation. The total cost of this error was $125 million. Trying to cope with a lack of a common definition for estimating the size of a GILT project can lead to similar problems.

This is reminiscent of the situation for general measurements before the advent of the French Revolution. A French foot ('pied du roi' - 12·79 inches) was different from an English foot that was different from the Welsh foot (9 inches). The basis of the current Imperial linear measures were unified by Edward I in 1308 who ordained (in a highly scientific manner for the 14th century) that an inch was to be three grains of barley, dry and round, taken from the middle of the ear and that twelve inches were to make a foot. It took the French Revolution to provide a (mostly) logical approach to establishing general units of measure based on a decimal scale (although somehow, the 10 day-week did not catch on).

GILT Metrics is a proposed LISA standard, aimed at providing a unified and verifiable (and unlike the French Revolution, a bloodless) way of establishing the size of a given localization task for electronic files. Why *metrics*? The *American Heritage® Dictionary of the English Language* (Fourth Edition) defines the noun *metric* as *A standard of measurement*.

Words and Characters

GILT Metrics mandates both word and character counts. Character counts convey the most precise definition of a translation task, whereas word counts are the most commonly used metric in the translation industry. GILT Metrics encompasses both measurements, thus affording both translation suppliers and customers with a choice as to which measurement most adequately reflects the translation task in question.

Canonical Form

One of the main problems with calculating word and character counts is the plethora of differing proprietary file formats that can contain a mix of form and content data. Trying to establish a standard that addresses all of these formats is impossible – the word count dragon has too many heads to attempt to cut them all off with one swipe. As soon as one head is cut off, a new one will appear somewhere else. A better approach is to force the dragon to enter a narrow passage where the heads are all forced together. Enter the XLIFF knight in shining armor called *Unicode*.

XLIFF is the OASIS standard for *XML Localization Interchange File Format* and is designed as a way of exchanging translatable data in an XML format. The GILT Metrics proposal relies on using the XLIFF representation as the canonical form for establishing the basis of word and character counts. The proposal mandates that all characters be counted in their Unicode representation and that all multiple space characters be reduced to a single character. In addition, word boundaries are defined with reference to *Unicode Technical Report 29 – Word Boundaries*. This provides an unambiguous definition of what constitutes a word.

By using XLIFF as the canonical form for counting the source language text, the GILT Metrics proposal establishes a common and well-defined format for word and character counts.

Example:
```
<source>An example of the canonical form of a text unit.</source>
```

Within, XLIFF inline codes are interpreted as inline XML elements. The inline elements are not included in the word and character counts, but form a separate inline element count of their own. The frequency of inline elements can have an impact on the translation workload, so a separate count is useful when sizing a job.

Example:
```
<source>In this <g id="g1">example</g> the in-line codes do not
feature in the word and character counts.</source>

<source>In this <g id="g1">exa<x id="x1"/>mple</g> the in-line codes
do not feature in the word and character counts.</source>
```

Standalone punctuation characters are also featured as an additional category in both word and character counts. They are included in the main count, but can be deducted from both by mutual consent if they do not increase the translation workload.

GILT Metrics addresses all issues related to counting words and characters in the XLIFF canonical format. Since the sentence is the commonly accepted atomic unit for translation, it proposes sentence-level granularity for counting purposes within XLIFF.

GILT Metrics does not preclude producing metrics directly from non-XLIFF format files, as long as the format for counting is based on the XLIFF canonical form for each text unit being counted. This can be done dynamically on the fly. In these instances, an audit file will be necessary for verification purposes.

In summary, the main goal of GILT Metrics is to provide a detailed count for words and characters based on the characteristics of individual sentences. The aim is to provide sufficient detail to enable an accurate definition of the scale of the translation task. The customer and supplier can then decide which of the statistics to use or not when costing the translation task for a given file.

Logographic Scripts

Word counts have little relevance for Chinese, Japanese and Korean source text. For these languages GILT Metrics recommends using only character counts.

Quantitative and Qualitative Measurements

GILT Metrics fall into two categories – how many, and what type. The primary count will always be unqualified, i.e., how many characters and words are in the file. This is the minimal conformance level proposed for GILT Metrics.

A typical translatable document will contain a variety of text elements. Some of these elements will contain non-translatable text, some will have been matched from translation memory and some will have been fuzzy matched by the customer. It is therefore important to be able to categorize the word and character counts according to type in order to provide a figure in words and characters for the GILT task.

Count Categories

GILT Metrics recommends the following count categories:
- Exact Matched Count – an accumulation of the word and character count for text units that have been matched unambiguously with a prior translation and that require no translator input.
- Leveraged Matched Count – an accumulation of the word and character count for text units that have been matched against a leveraged translation memory database.
- Fuzzy Matched Count – an accumulation of the word and character count for text units that have been fuzzy matched against a leveraged translation memory database.
- Alphanumeric-Only Text Unit Count – an accumulation of the word and character count for text units that have been identified as containing only alphanumeric words.
- Numeric-Only Text Unit Count – an accumulation of the word and character count for text units that have been identified as containing only numeric words.

- Punctuation-Only Text Unit Count – an accumulation of the word and character count for text units that have been identified as containing only punctuation.
- Standalone Punctuation Count – an accumulation of the standalone punctuation word and character counts from the individual text units that make up a document.
- Measurement-Only Count – an accumulation of the word and character count from measurement-only text units.
- Other Non-Translatable Word Count – other non-translatable word and character counts.

Verifiability

Any measurement standard must have a reference implementation as well as an authoritative body that tests and validates the measuring instruments. In the USA, this is provided by the National Institute of Standards and Technology. In order to be successful, GILT Metrics must provide for a certification authority that will (1) maintain reference documents with known metrics and (2) provide an online facility to test given XLIFF documents. In this way, both customers and suppliers can be safe in the knowledge that GILT Metrics provides an unambiguous and reliable way of quantifying a GILT task.

Summary

The GILT Metrics proposal is based on well-defined standards:
1. XLIFF
2. Unicode ISO 10646
3. Unicode TR29

GILT Metrics proposes maintaining counts for words and characters, standalone punctuation and inline code and references. It also recommends additional qualitative counts for the text element categories detailed above. All of this detail allows a precise and unambiguous definition of the GILT task for a given electronic file. This rich detail allows suppliers and customers to be able to precisely measure the task at hand and to more easily do business with one another as a greater level of trust is generated.