

# Opening the Black Box, Part I

Demystifying the Processes of  
Localization and Translation

By

**John White**

[www.ventajamarketing.com](http://www.ventajamarketing.com)

- I already understand localization. Or do I?
- What do translators really do with my product?
- Why is localization so expensive?



## Introduction

“Why is localization so expensive?!”

We hear this question a lot from our clients, and at one time we had a short answer for it:

Words.

Our clients usually just scowled at us when we gave that answer, and so we elaborated a bit:

The expense-perspective: You paid to create your English-language product, but because your engineers and writers use English words, it looked to you as though you *didn't* pay anything to create it. Now you need to write a fat check to somebody in order to create other versions, and you're annoyed because “all they're doing is translating,” which feels like child's play compared to the work you've done.

The revenue-perspective: Your investment in the English-language product will be returned by lots and lots of English-speaking people who will give you money because you solved their problems. Similarly, localization is an investment in a [German/Japanese/Korean/Russian/French/...] product, and this investment will be returned by lots and lots of non-English-speaking people who will give you money for solving *their* problems.

In other words, there is an expense-side and a revenue-side to the coin of localization.

We also hear that many people in technology consider the process of delivering international products a “black box”. Accordingly, Part I of this paper explains the terms and steps in “Internationalization” and “Localization,” with a few grisly details that are second nature to translation professionals, but which look like a black box to most of their clients. In Part II we describe where high localization costs come from, and what organizations can do about the side of the black box over which they do have some control. (When all is said and done, it still comes down to the first answer: Words.)

## Background: Internationalization

It is best in the long run to first *internationalize* the software, as in the Portuguese example in Figure 1, so that no matter how it changes for user interface or business logic, the code base at the core of the product is always the same (sometimes called a *single worldwide binary*). To *externalize* those features and characteristics that can change from one region (or *locale*) to another—e.g., language, color scheme and accounting standards in software, or narrative examples, conditional text and legal text

in documentation—is to place them in separate *resource files* that vary from locale to locale, then call them from a single, common code base.

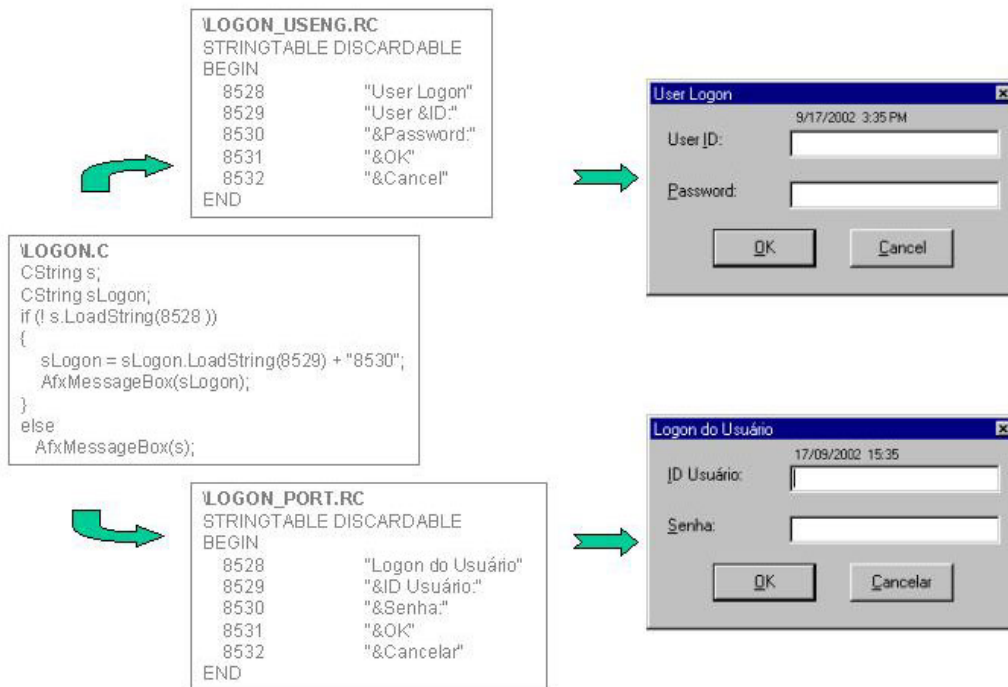


Figure 1

This process of internationalization (or I18n, because there are 18 letters between the “I” and the “n”) makes life easier for a number of people along the value chain.

I18n helps:	Because:
Engineers	Cosmetic aspects of code base require less maintenance.
Release Engineers, Configuration Managers	Different language-versions are all built with the same code base. There is only one code base to patch/develop.
Translators	They are able to stay away from code.
Project Managers	There are fewer surprises and incremental handoffs of newly discovered text for translation.
Support Technicians, Users	It eliminates the chance of English text popping up at embarrassing moments, especially in error messages.

There are many other aspects of I18n – Unicode, double- and multi-byte character set enabling, overall software architecture – but ultimately, companies internationalize their products not only because it’s better engineering, but also because they can reduce time to market and make/save more money in *all* regions.

Of course, it’s nearly impossible to get I18n right the first time, and it usually requires multiple product cycles to refine the process, so the enlightened organization takes a

long, patient view. Impatient organizations may leave locale-specific elements, such as strings or error messages, inside the code base, usually to get the English-language product to market in a hurry. This is not evil, but there are costs associated with it later in the process.

Companies which outsource I18n effectively leave much of the process inside the black box. For those companies whose own engineers and writers internationalize their products, I18n is not so mysterious.

## Background: Localization

Few companies, however, *localize* their products themselves, because few companies can justify the cost of keeping translators and localization engineers on staff.

Localization (or L10n) is the process of delivering a product that meets the needs of users in a specific locale, and because most companies outsource this function, it usually lands squarely in the black box.

The L10n process is not completely different from or contrary to the process of creating most technology deliverables. When the product has been properly internationalized, L10n is a parallel function that takes a copy of the resources from their normal flow in development, modifies the copy for the needs of a specific locale, and joins the original further downstream in the development flow (Figure 2).

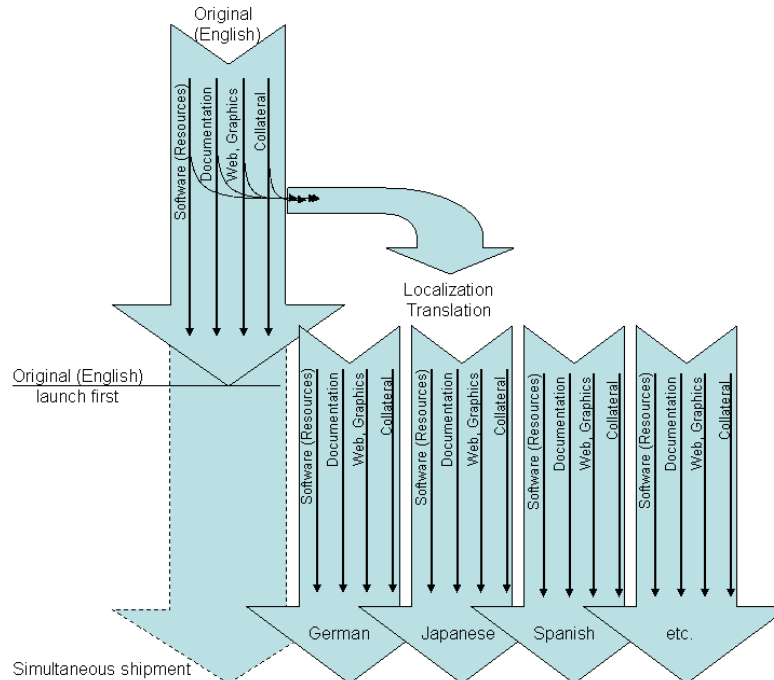


Figure 2

There are a few localization concepts, however, which are alien to most people who build technology products, and these concepts are inside the black box.

## Key Concepts

### Glossary (Terminology List)

To ensure uniformity of translation throughout the product (and, as the international effort grows, throughout the company), it is a good practice to put in place a glossary, which contains approved translations of key words and phrases. A translation glossary gives the *equivalent* of the key terms in the target language.

English	Japanese	English explanation (do <u>not</u> translate)
churn	客離れ	i.e., turnover among customers
Class	クラス	A class is a template that defines the generic characteristics of an object or module in the system, that is, a file.
Crunchware	Crunchware	Noxlate; this is the name of client's company and client's product
Transmission Control Protocol/Internet Protocol	Transmission Control Protocol/Internet Protocol (TCP/IP)	Noxlate
Transmit Speed	転送スピード	"Transmit" is not a verb here

Figure 3

The Explanation column in Figure 3 is very important for preserving contextual information for the benefit of the translators. Note also that the glossary plays the important role of dictating what should *not* be translated.

Here are some key moments in the life of a glossary:

1. Client hands off early version of product to localization vendor for creation of glossary.
2. Localization vendor compiles list of key terms, with contextual comments.
3. Client conducts training session for translators and editors (optional, and too often overlooked)
4. Translator translates (or, in some cases, doesn't translate) into target-language equivalents.
5. Vendor returns glossary draft to client.
6. Client sends glossary out for review by stakeholders most likely to complain about undesirable translations, in order to avoid these complaints once the product has been released. (This is extremely important, and should be performed by in-country partners and co-workers whose livelihood depends on the quality of the translation.)
7. Client returns glossary comments to vendor, who incorporates them.

8. Once approved, the glossary goes to translators, reviewers, editors *and client* for continued reference.

A typical glossary will contain a few dozen up to a few hundred terms.

### **Simultaneous shipment (Sim-ship)**

Handing off resources for translation early in development allows the localization process to begin sooner, but the original resources are more likely to change and grow, requiring additional handoffs before release of the original. On the other hand, handing off the resources after the release of the original ensures that they are frozen, but delays the launch into the regional markets (see Figure 2).

Simultaneous shipment in multiple languages is attainable, but it usually takes several localization cycles, as well as ironclad buy-in from upper management. The localization process needs to be an integral part of mainstream development so that changes to the original move quickly into the localized versions.

### **Machine Translation (MT)**

The urge to automate as much of human effort as possible has also touched the specialty of translation. Since the 1950's, the field of computational linguistics has contributed a great deal to the technology behind computerized translation, and the road ahead is filled with promise.

In MT, the computer applies rules and algorithms for syntax, morphology, semantics and other rules to translate text into a destination language. Another approach uses statistical models to arrive at the most likely translation for the input text. Depending on the source-destination language-pair (some are better matched than others), the unedited result will almost always preserve meaning, but it will rarely be as natural as if translated by a human native speaker.

Some localization vendors use MT as adjunct technology in their translation workflow. It can save money and time as an interim step to a post-translation editing pass, but by and large, human language is too old and MT not (yet) old enough for most clients to entrust completely the localization of, say, technical documentation entirely to a computer.

### **Translation Memory (TM) and Computer-Assisted Translation (CAT) Tools**

While also computer-based, these tools differ from Machine Translation. The goal of MT is for the computer to bear the brunt of the translation work, whereas TM and CAT tools help the translator do his/her own translating better, more accurately and faster. Localization vendors in the 21<sup>st</sup> century *must* use these tools not only to compete on price, but also to meet market expectations of consistency and quality of translation. Currently prominent products in the TM/CAT category include Catalyst, Déjà Vu, Passolo, SDLX and Trados, among many others.

TM expands on the idea of the glossary. Beyond correlating a few dozen or hundred key terms, TM creates a one-to-one correlation between *all* of the source text and *all* of the destination text in the entire product and places these correlated pairs into a scalable database. Figure 4 shows source text on the left and destination text on the right.

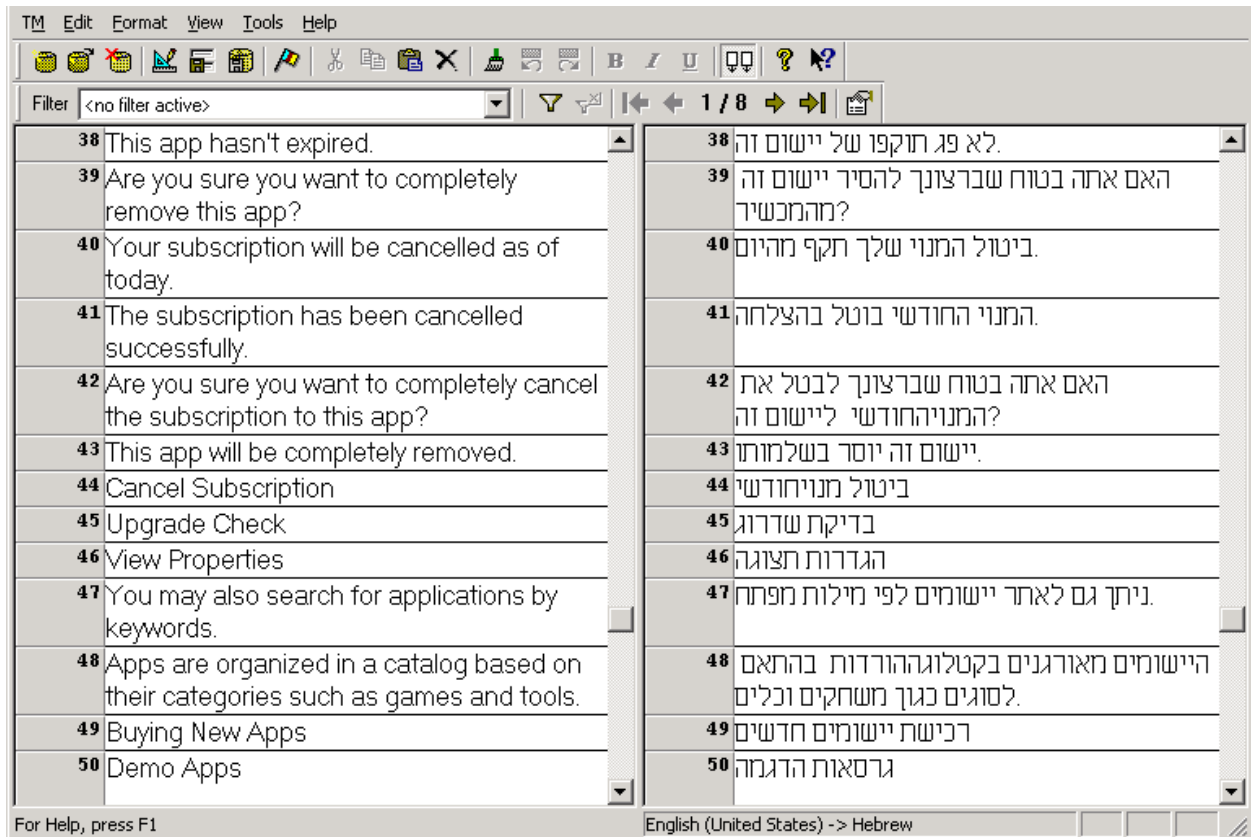


Figure 4

The database may contain text from software resources, documentation, Web pages and marketing collateral, making all of it available as reference material to any translator working on any of these projects. In addition, fuzzy-matching algorithms rate approximate matches, so if the TM software finds similarity between a new sentence and another sentence already translated in the database, it will suggest it to the translator with a percentage-rating of closeness.

## Benefits of TM

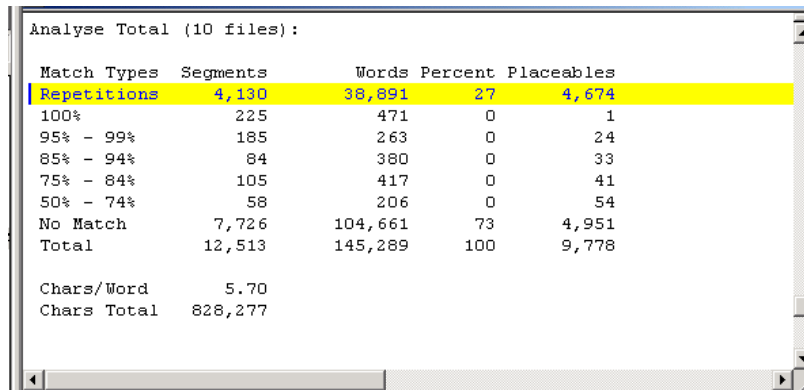
Several compelling benefits can accrue to the organization with all (or even most) of its translation memory in a database.

### ***Analysis and cost estimates are more accurate***

As observed somewhat ironically on page 2, the key metric in the cost of a localization project is Words. Before TM, it sufficed to estimate the wordcount of the entire project and multiply it by a price per word, but why pay to re-translate text that has already



been translated, or which appears identically in many different places? With TM, it is possible to determine the wordcount of phrases and sentences that have already been translated, and thereby arrive at a more accurate cost estimate.



Match Types	Segments	Words	Percent	Placeables
<b>Repetitions</b>	<b>4,130</b>	<b>38,891</b>	<b>27</b>	<b>4,674</b>
100%	225	471	0	1
95% - 99%	185	263	0	24
85% - 94%	84	380	0	33
75% - 84%	105	417	0	41
50% - 74%	58	206	0	54
No Match	7,726	104,661	73	4,951
<b>Total</b>	<b>12,513</b>	<b>145,289</b>	<b>100</b>	<b>9,778</b>
Chars/Word	5.70			
Chars Total	828,277			

Figure 5

Figure 5 shows the analysis on a batch of ten new files submitted to a vendor for translation.

- 38,891 words are in 4,130 repeated segments<sup>1</sup>, or segments which exactly match another segment in the ten files. The translator can translate the first occurrence of these segments and the TM software will propagate them throughout the project. In addition, 4,674 words are “placeable” (numbers, tags, symbols) and do not require translation.
- In descending buckets from 100% down to 50%, there are a few hundred words in segments which have been translated before. These buckets represent descending degrees of fuzzy match.
- The software finds no match for segments containing 104,661 words, so these must be translated from scratch. The 4,951 placeable words do not require translation.

Many vendors offer discounts based on this analysis: The higher the percentage match bucket, the greater the discount on words in that bucket.

- A 100% match means, of course, that no translation work is required, but the words in the 100% segments must still be “touched” (engineering, desktop publishing, translation memory work, final review, QA), so few vendors discount them entirely.
- A 95-99% match often means that punctuation or the spelling of a single word has changed, or a word has been added or deleted, so a translator must do

<sup>1</sup> A *segment* is the lowest level of granularity that the TM database holds. Broadly, and depending on the segmentation rules of the TM software, it is usually any text (words, phrases, sentences) followed by a hard return, tab or sentence-ending punctuation.



some light work on the segment and the vendor will discount the segment slightly less.

- Below 75% matches, however, discounts are less common because, by the time the translator has found and dealt with the differences between the old and the new text, s/he may as well translate it from scratch.

### ***The vendor can pre-translate new versions***

Before the new file has made it as far as the translator, the TM software will have pre-translated as many 100% matches as possible. For any segment that already has a match in the TM database, the software will retrieve and place the corresponding translation. This greatly reduces the translator’s work and shortens the time to deliver the completed product.

In Figure 6, the Chinese segments marked with gold (lines 164-171 in the right-hand column) are 100% matches for the corresponding English segments on the left, which have been retrieved from TM and dropped into the translator’s work file in advance.

164	File Memory Full.	164	文件内存已满。
165	To make enough storage space to install this app, the following app(s) will be temporarily disabled: %s.	165	为了腾出足够的存储空间以安装此软件, 将暂时禁用以下软件: %s.
166	You can restore disabled apps by simply starting them, at no additional purchase cost.	166	您只需启动这些软件即可恢复禁用的软件且不必另外付费。
167	Proceed?	167	是否继续?
168	[Size: %s]	168	[大小]: %s]
169	Connecting...	169	正在连接...
170	Cancelling...	170	取消...
171	Unable to connect to the wireless network.	171	无法连接到无线网络。

For Help, press F1      English (United States) -> Chinese (PRC)

**Figure 6**

### ***Translations enjoy leverage from one version to the next***

Similarly, the software identifies fuzzy matches and places them. The translator modifies the existing translation in light of how the source text has changed and adds the new segment to the TM database. This helps the translator spot English segments that have changed since the last time the product was translated.

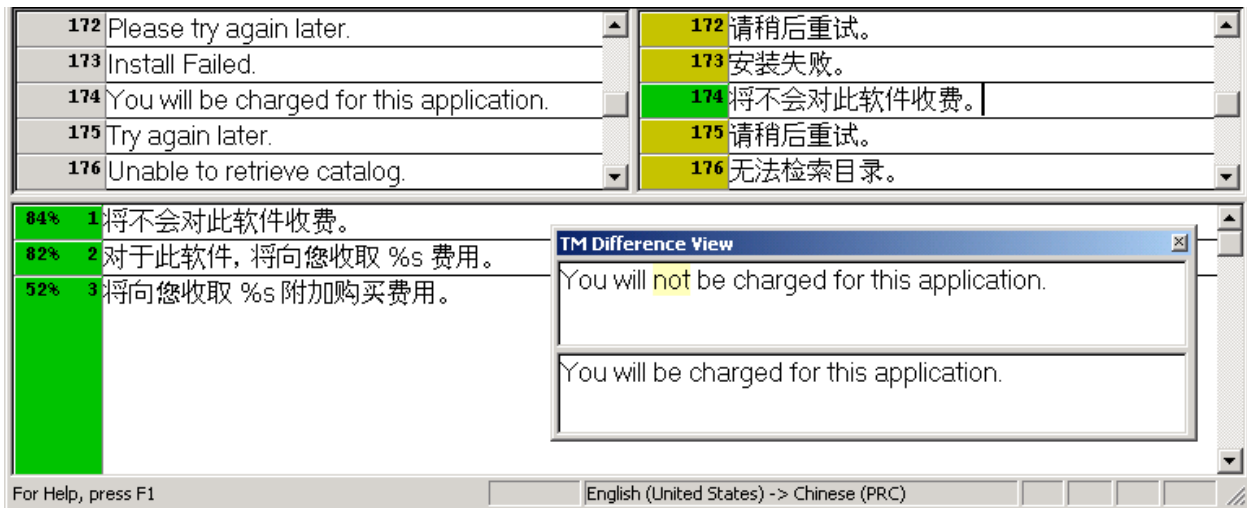


Figure 7

In Figure 7, segment 174 reads “You will be charged for this application.” The software found the closest fuzzy match (84%) and pre-translated it, tagging it green to call the translator’s attention to it. The TM Difference View window shows the very important change made to the English sentence since the last round of translation: The sentence now reads “You will **not** be charged for this application.” The software provides enough of the original translation so that the translator does not need to start from scratch.

The leverage from one version of the localized product to the next is a tremendous advantage of TM software. While it does not lower the overall wordcount of a product, it eliminates needless work for the translators and shortens time to market for the localized versions.

***All terms are available for lookup and concordance search***

The combination of the TM database and fuzzy matching also allows for concordance searches on specific text for similar, but not necessarily identical, occurrences.

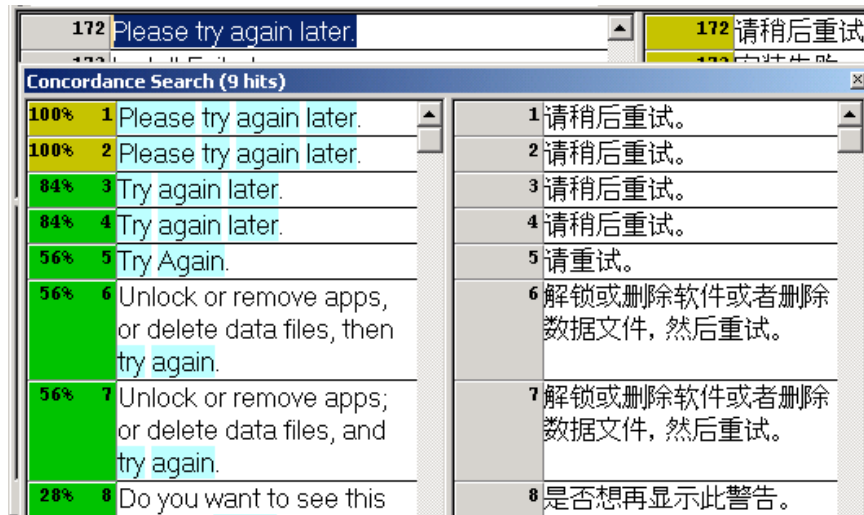


Figure 8

In Figure 8 the translator has looked up “Please try again later” in the TM database to see in how many different ways similar text has been translated in the past. This functionality goes deeper than that of a translation glossary because it broadens the subject of the lookup from key terms to common phrases, and the domain of lookup from the few hundred glossary terms to the entire TM database.

### ***Clients preserve history from one vendor to the next***

Finally, the TM database represents a valuable asset if/when the time comes to change localization vendors. With TM, if a vendor goes out of business or is unable to scale to meet a client’s localization needs, the client can forward the TM database to the new vendor, who can then exploit the translation history with less delay.

There may be technical limitations (different TM database formats) and legal issues (ownership of the TM database itself), but the larger the translation history, the smaller these issues look by comparison.

### **A few more notes on TM**

- The real value in TM lies in its continued use *over time*. A sustained international effort to deliver future versions of localized products will benefit handsomely from TM. The organization interested in a one-time, quick-and-dirty translation will enjoy far fewer benefits.
- Although TM saves a lot of work, it also involves a lot of work for vendors. Some vendors bill for it directly and conspicuously, while others bill for it under general engineering costs. On balance, though, its benefits outweigh its costs, and there is almost no point in trying to save money by instructing the vendor not to use TM tools.

- The client must bear in mind that translation memory tools are *not* machine translation tools. As described above, MT tries to calculate translations of new text using rules and existing translations, whereas TM accumulates segment-by-segment history and assists human translators. While hybrid TM-MT solutions are becoming more popular, the industry is still a long way from reducing translation to a pure matter of software and hardware.
- The client must bear in mind the concept of the *segment*. TM looks for and matches text in entire segments because this is the lowest level of granularity which the software can use. For this same reason, wordcounts are a function of words in segments, as described in Figure 5.

Clients who do not understand the concept of the segment ask a very common question: “My company and product name, Crunchware, appear 900 times in the product, and I want them preserved as “Crunchware” in all languages. Do I have to pay for that?” The answer is “Sometimes.” If “Crunchware” occurs as a segment all by itself in the TM database – and it probably does – then at all occurrences of that same segment there is no work for the translator to do, and the client will likely receive a steep discount for that 100% matched segment in every place that “Crunchware” appears alone. However, if “Crunchware” appears in a completely new sentence with 44 other words – like this sentence, for instance – then the software will report a 45-word segment with no match in TM, and the vendor will likely charge for 45 words at the full rate per word.

## Summary

A properly internationalized product is a delight to localize, because it involves no wasted effort in handoff, cost estimation, scheduling, translation, rebuilding, testing, release or support. It meets the needs of users in other regions with no changes to its core functionality, no patches, no bug lists, and no excuses. Enlightened organizations manage their own internationalization (I18n), which keeps it out of the black box.

The black box looks black mostly because localization (L10n) takes place outside of the client’s organization and involves tools and skills rarely found inside the organization. The black box is not so much product *development* as product *transformation*, in which a familiar, English-original product becomes a French/Korean/Russian/Hebrew/... copy. Translation memory (TM) tools are at the heart of the transformation, keeping costs as low as possible and accelerating the work of translators.

Although the L10n process seems trivial to many technology clients, there really is a great deal more to it than simply translating words from one language to another. Honest.

## Next Steps

You've learned something from this paper, haven't you? You'd like a strong process, plan and player for your localization project, wouldn't you?

To give your localization effort every chance of succeeding:

1. Help yourself to other resources on our Web site.
2. Become as conversant in localization terminology as you can.
3. Contact us for a **free assessment** of your project, before you paint yourself into any corners.

John White of venTAJA Marketing ([johnw@ventajaNOSPAMmarketing.com](mailto:johnw@ventajaNOSPAMmarketing.com)) offers localization project management and international product management for technology companies, and has managed internationalization and localization projects since 1992. He sometimes wakes up in the middle of the night dreaming in segments.

Copyright 2007, venTAJA, Inc. Trademarks and brands are the property of their respective holders.