Do TM Systems Compare Apples and Pears?

Philippe Mercier

What are the difficulties or potential problems that may arise if we want to switch from one TM software application to another? Can we exchange TMs between systems; are they compatible? Can we switch from one system to another? What will be the cost of such a move? Is TMX the solution?

Likewise, can we use the same TM when translating an XML document, an MS-Word document or a FrameMaker document? We can, but some unexpected results may occur.

Why should we worry about this?

The problems that may arise are quite simple: segments that are considered as 100% matches by one TM Software application could be considered as lower matches by another software application. In other words, 100% matches in one TM software application could become 97% matches, 70% matches or maybe even less in others. The same could happen when changing the documentation format. So, if we have millions of words being calculated as 100% matches in our TMs, we obviously don't want to lose a high percentage of these.

What Is a Translation Memory System?

To understand how this could happen, IF it happens, and how we can avoid it, we must first understand what a translation memory system is and how it works.

A Translation Memory, or TM, is a repository of human translations. It is filled by a TM software application which stores each sentence and its associated translation into the memory as soon as it is translated. The elements stored in that repository are called translation units or TUs (See Fig. 1).

When a new sentence needs to be translated, the translation memory software application can first run a search to see whether or not that sentence already exists in the translation memory. If it is found, the associated translation will be proposed to the translator. When searching the memory, three situations can happen:

Nothing is found. This is called a *no match* in TM terminology. The translator has to translate the full sentence.

The exact same sentence is found. This is called a *100% match*. The system will show the translation to the translator who will, most of the time, simply validate this translation in the current context.

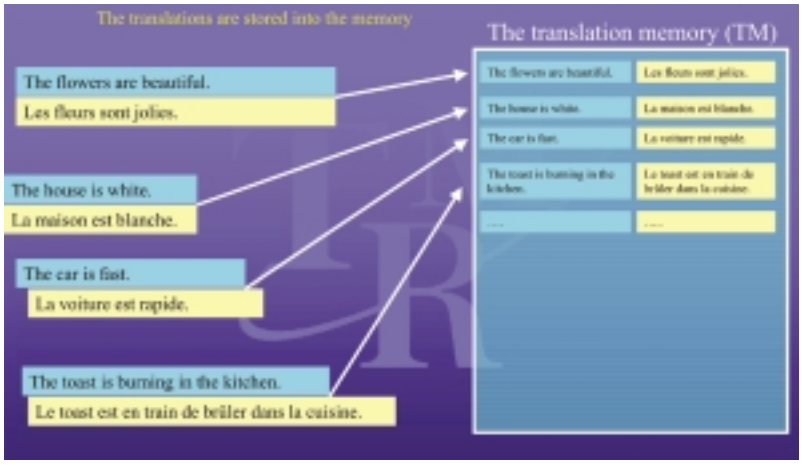


Fig. 1: Each translation unit stored into the TM is composed of two elements: a source segment and its translation.

A similar sentence is found. This is called a *fuzzy match*. Depending on the differences between the source sentence and the sentence found (number of different words, position of the words), the system will calculate a similarity match and express it as a percentage. We'll say that this is a *75% fuzzy match* or an *84% fuzzy match* or any other value. The translator can then modify or adapt the proposed translation.

Original Sentence	Segmentation done by TM Software "A"	Comment	Segmentation done by TM Software "B"	Comment
Harris received a Ph.D. in Political Science from the University of Wisconsin-Madison.	Harris received a Ph.D. in Political Science from the University of Wisconsin-Madison.	"Ph.D." has been recognised as abbreviation	Harris received a Ph.D. in Political Science from the University of Wisconsin-Madison.	"Ph.D." has NOT been recognised as abbreviation

Fig. 2: Comparing two existing TM software applications

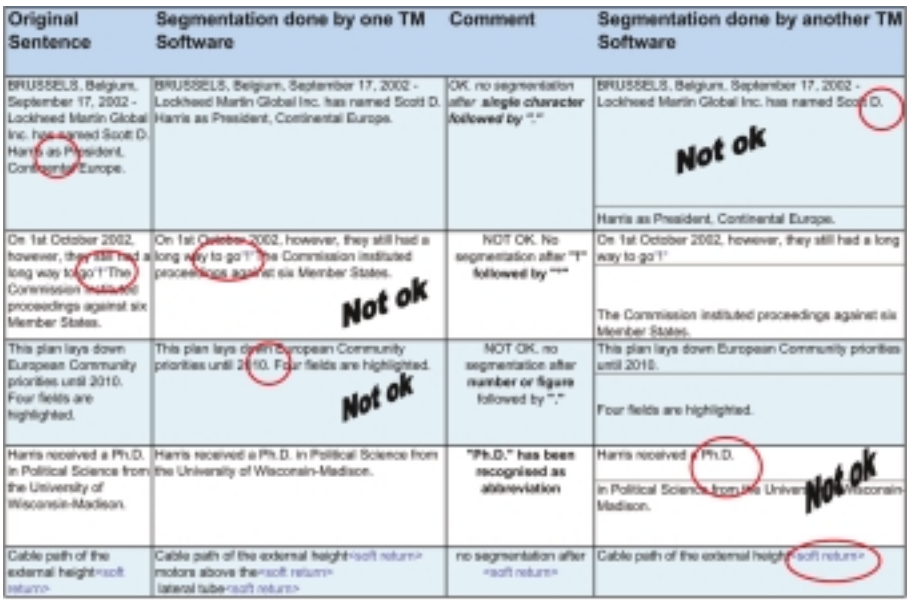


Fig. 3: Examples of segmentation algorithm differences of main TM software applications

The Translation Process

Now, let's see how this works practically for the translator. Many different TM software applications are available on the market today but they all work the same way. The translation process is always the following:

Step 1 - Break the document into segments. A segment is generally a full sentence but it can also be part of a sentence or even a full paragraph. It is the base unit that will be translated and stored into the TM along with its translation.

Step 2 - Check if a segment, equal or similar to the current segment, is already in the memory.

Step 3 - Depending on what has been found (or not found) in the TM, the translator can accept the proposed translation or type in a new one.

Step 4 - The source segment and the translation are recorded into the TM.

Step 5 - The system goes to the next segment and the cycle starts again at step 2. This cycle continues until the last segment of the document is translated.

We Must Compare Apples and Apples

When searching for a potential translation in the translation memory, it is critical for the TM system to compare apples and apples, and not apples and pears.

A TM system stores and searches for « segments » in a translation memory but what exactly is a segment? Is it a sentence? Is it a full paragraph? Is this important? Of course it is. The way to define a segment is essential. We must search the TM for the same kind of data that has been stored into that TM before: search sentences if we have stored sentences; search paragraphs if we have stored paragraphs, etc.

For example, if we decide that a segment is a sentence, then the translation memory will contains sentences and their translations. If later we change the definition and decide that a segment is a paragraph (which can of course contain several sentences), we'll search for the translation of a full paragraph by comparing the current paragraph with all the sentences existing in the translation memory: we compare apples and pears. Obviously, we'll never find a paragraph that is equal to a sentence, or only extremely rarely (when a paragraph contains just one sentence). So, if we change the segmentation between two projects in any substantial way, we'll probably lose ALL our translations and the TM will become useless.

This is the worst case scenario and should not happen of course, but it serves to effectively illustrate the problem.

A similar situation can happen if we build a TM with one TM software application and then later decide to use it with another software application. The way each software carries out the segmentation may be slightly different and that will cause some problems similar to the ones we have just described.

But those problems could also happen when using the same software, as we'll see below.

Segmentation Rules

The method used by a TM software application to break the text into segments is called segmentation. Segmentation is based on rules; a rule tells the system what to consider as an end of segment marker.

Segmentation, more often than not, consists in breaking the text into sentences. Thus, the most common rule defines the '.' as an end of segment marker. This means that the system will start comparing each character of the document and when it encounters a '.', it will decide that the end of the current segment has been found (See Step 1 of the process defined above). That segment will then be presented to the translator for translation (steps 2 & 3).

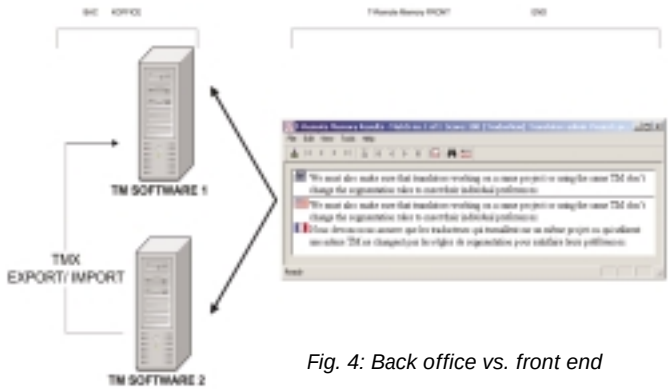


Fig. 4: Back office vs. front end

To understand this, let's take the paragraph at the beginning of this section, above in this article as an example. With default rules, a TM system with only the '.' rule will consider that it contains two segments:

Segment 1: « The method used by a TM software application to break the text into segments is called segmentation. »

Segment 2: « Segmentation is based on rules ; a rule tells the system what to consider as an end of segment marker. »

Those two segments will be translated and stored into the TM.

Now, let's suppose that six months later, we need to translate an update of this document. A new rule has been added and defines the ' ;' as an end of segment marker. Our paragraph above will thus now contain three segments instead of two:

Segment 1: « The method used by a TM software application to break the text into segments is called segmentation. »

Segment 2: « Segmentation is based on rules; »

Segment 3: « a rule tells the system what to consider as an end of segment marker. »

When searching for the translation of segment 1 in the TM, the TM system will of course find and retrieve it from the previous job as a 100% match. No problem. But, when searching for the translation of segment 2, it will never find anything (too many differences with previous segment 2: considerably fewer words now). And finally, when searching for a potential translation of segment 3, it won't find anything at all either, for exactly the same reason.

Let's do some calculations now. If our document contains only those three sentences, it has a total of 37 words. Because we have changed the segmentation rules (or switched to another program that has different rules), only 17 words (segment 1) remain as 100% matches while we have to retranslate 20 words from scratch (we know, however, that the translation is in the TM but the system won't find it because segment 2 is now divided into two segments!). We can see that more than 50% of the total words need to be retranslated because of the different segmentation rules. If we extrapolate that to documents containing millions of words, the cost of having changed the segmentation rules could become astronomical.

We must also make sure that translators working on a same project or using the same TM don't change the segmentation rules to meet their individual preferences. If this were the case, the TM could become a real mess ; it would get filed with completely different segments, and with consequences that we already know (i.e. an extremely low reusability of the segments).

So, we strongly advise NOT to change the segmentation rules and, if you decide to switch to another TM software application, make sure to configure it with the same rules that you have been using before instead of using default rules.

Note: Some other common end of segment markers are the end of paragraph sign for example the '!', the ' ?' and the tabulation character.

Abbreviation Lists

Another important part of a segmentation algorithm is the abbreviation list. As we have seen, the '.' is the most common segmentation rule. But it is also the character used in abbreviations such as 'Mr.', 'Dr.', 'etc.' and many others. Thus, when abbreviations are used in the middle of a sentence, there must be a way to tell the system that the '.' should not be considered as an end of sentence marker. To achieve this, the segmentation algorithm will consult its abbreviation list and if the word followed by the '.' is in the list, it will be ignored and won't be considered as an end of sentence marker.

So, this abbreviation list influences the way a program carries out segmentation. In Fig. 2, we can see an example of comparing two existing TM software applications. In TM Software A, the abbreviation 'Ph.D.' is defined and the system does not segment on the '.' on 'Ph.D.'. On the contrary, the abbreviation is not defined in the list of TM Software B, and the system considers the '.' of 'Ph.D.' as an end of segment marker, which causes a segmentation error as we can see (the system detects the presence of two segments instead of one).

Thus, if the TM has been created with software "A", the full sentence will be translated as a segment and stored into a TM. Later, if we move to software "B" or change the list, the system will search for the translation of "Harris received his Ph.D." as a segment that it will never find because too many words are different. Then it will search "in Political Science…." as a segment also, which it will never find either.

The conclusion is the same as above: don't change the abbreviation list and if you switch to another program, discard their list and use yours or you may lose a lot of your TMs.

Other Cases

There are some other cases of dots that the system should not consider as end of segment markers.

Here are some examples:

- *The house of G.W. Bush is white.*; initials of names
- *Article 7. Minimum*; dots after numbers
- *www.telelingua.com*; dots in url's
- *My phone number is +32 2/373.68 .68.*; dots within numbers
- *B. European environmental policy*; dots in enumerations.

And here, there is nothing we can really do to influence the segmentation algorithm. Some TM systems will handle these cases properly, some will not. This will imply some minor differences in the segmentation and thus incur some losses of 100% matches, but these should be minimal as we don't have urls or phone numbers on every line.

In Fig. 3 we can see more examples highlighting the segmentation algorithm differences of the main TM software applications available on the market today (using default rules):

We see also that some bugs (clearly the software should not segment on the '.' of Scott D. Harris») can cause segmentation differences and again this is something that cannot be fine-tuned.

Back Office vs. Front End

A very simple solution to avoid all segmentation problems between different TM systems would be to separate the TM handling from the segmentation handling: one software does the segmentation on the translator computer and the TM Software applications only handle the management of the translation memory (storing segments, searching segments).

For example, if you use T-Remote Memory (a software developed by Telelingua that allows the use of Translation Memories remotely via the internet) as a front-end (translator interface) and any other TM program in the backoffice as a translation memory engine, you eliminate all the segmentation problems described above instantly. You can exchange memories or switch to another system at any time (See Fig. 4).

Why? Simply because, as we mentioned earlier, the segmentation algorithm determines what is going to be the base information unit stored into the TM but also defines the information that will be searched for in the TM. If T-Remote Memory has a ';' segmentation rule, only segments following that rule will be stored in the translation memory handled by TM Software A for example.

If at a later point in time, the TM created with TM software A is exported and imported into TM Software B, and the translators start working on a project, the segmentation won't change if they are also using T-Remote Memory as an interface. And thus the backoffice TM software B will be asked to search in the TM for the exact same segment that was stored by TM software A, which it will of course find.

What About Formatting?

Formatting is another serious problem. Depending on the format of the original document (XML, HTML, MS-Word, Framemaker,…), a TM created when translating one kind of document won't be 100% usable when translating another kind of document.

Why not? Well, simply because the internal software codes for bold, italics or any other formatting won't be the same.

For example, if we were to translate the sentence: « The toast is burning in the kitchen » from an MS-Word document and from an HTML page, here are the two results that would be stored in the TM (pay attention to the codes recording the fact that the word «toast» is in bold):

- HTML: The {\cs6\f1\ cf6\lang1024<B>}toast{\ cs6\f1\cf6\lang1024 </B>} is burning in the kitchen.

- MS-Word: The {\b toast} is burning in the kitchen.

As you can see, the HTML formatting of bold «{\cs6\f1\cf6\lang1024 <B>} » is not the same as the MS-Word formatting of bold «{\b».

As such, if we have a TM created from MS-Word documentation and if we move this documentation to HTML or XML format (or any other), the system will find exact matches for sentences in the new HTML or XML format but it will see that the formatting codes are different and will consequently apply a formatting penalty to the segments found. You will thus lose all your 100% matches for all those segments containing formatting! The formatting penalty is very low but this is a real problem (see Fig. 5).

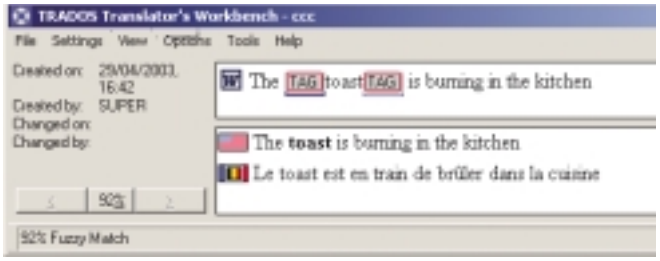


Fig. 5: The system applies a penalty difference and a 100% match becomes a 92% match.

So you might want to reconsider before moving all your documentation from MS-Word to XML for example, if you have millions of words in your TMs; otherwise all your 100% matches may become 92% matches.

TMX Format

The TMX format is an exchange format that allows you to export a translation memory from one system and import it into another system. However, one must be aware that this does not address the segmentation problem. It is merely a standard format used to exchange data between two TM systems. The Oscar TMX segmentation committee has some interesting ideas for incorporating segmentation rules inside the TMX itself, but this has not been developed yet.

One must also be extremely careful of the fact that there are two levels of TMX: TMX level 1 which does NOT support formatting and TMX level 2 which does support formatting. What does that mean? Well, TM software applications are supporting this TMX standard at various levels since developers are not always eager to offer a solution to export TMs and thus provide people with an opportunity to move away from their system to another. Some, for example, do not export the formatting properly, and it will then be totally lost when switching from one TM software to another.

The issue of formatting is extremely important and you should run extensive tests before deciding to make any move.

Conclusion

We see that many of the problems mentioned throughout this article can be avoided with a bit of finetuning and good preparation. If you don't want to lose your 100% matches, then don't change the segmentation rules between projects, make sure that all translators working on a project are using the same rules, and don't change the abbreviation list.

The same rules apply if you decide to switch to another TM software application. Use TMX to convert the memories but first do a test to see if the formatting is properly converted. Then, don't use the new software's default rules: make sure to

use the same segmentation rules and abbreviation list as before and everything should be ok.

You may lose a few 100 % matches because of the differences in segmentation with special cases but that shouldn't be critical or significant.

Finally, if you want to change the source format of your documentation, and move it from MS-Word to XML for example, you will lose a small percentage of matching for segments containing formatting (the formatting penalty percentage). While this percentage is low, unfortunately there is really nothing you can do to avoid this. ■

Philippe Mercier, the author of 10 computer programming books published in France, is the managing director of Telelingua Software SA. He can be reached at pmercier@telelingua.com.

A Little Paranoia Can Go a Long Way (I)

Roman Civin

Testing and bug-tracking is not an exact science, though most of us would like it to be. The more experienced we get as testers or test managers, the more aware we are of things lurking around corners, and the what-could-go-wrongs that will go wrong. Sounds a little paranoid, right? In this first instalment about what makes us shudder in localisation testing, we describe some elementary and more complex issues that testers encounter (or not) while providing some hints to help solve them. After reading, I trust you will be a little more paranoid too!

A healthy approach to testing should be tinged with a bit of paranoia. Add in a few parts of mistrust and you will be in good shape. Feeling bad vibes can support a good risk analysis, detect inappropriate test coverage while suggesting improvements, avoid redundant preparation, ameliorate an inadequate test plan, and for good measure, find leaks in the test process for a given product. Bearing this in mind, we must remain QA-conscious and flexible; otherwise we can become paralysed, instead of being an inconsolable worry-wart. Being "appropriately mistrustful" while adhering to a "customer is always right" attitude is a contradiction that must live together in the area of localisation testing. The following are several real-life examples. If you are a localisation testing expert, you may have demons such as these discussed here.

Test Preparation Control: Create, Then Choke?

Even some advanced projects have just a week of test preparation. You receive the key information and competencies only a few days before the start date. We must always be ready for this scenario and I will discuss this in one of the following issues of *Localisation Focus*. For the moment, let us focus on projects for customers who have a major product release once every year or two. Before the behemoth gets rolling, a lot of effort and trust is put into innovation of the cooperation model and the process or tools with the client's established partners. A dream, right?

What more could you wish for? You are invited to contribute, suggest, implement and you have time to adapt and optimise your own environment for the new big project. It adds great value and helps build a lasting relationship. Together, you decide to upgrade the test database by reporting features and allowing the partner team to test the new/fixed features. You cooperate on creating a tracking and reporting model; detailed and elaborate testing categories and a benchmarking model for testing volumes. Now imagine that upstream, you (the vendor) provided other complex services for the software publisher including translation, software engineering, etc. for this project. You are so ahead of yourself that the feedback and preperation stage can last a quarter.

The situation then becomes something like this: the testing stage typically would start much later than the localisation stage, yet we are encouraged to work on areas that should happen a few months later. As the testing start comes closer and the localisation kick-off meeting is well behind you, the then-promising testing initiatives receive less attention and are put aside unfinished as "extras", because things like product-readiness and tool key functionality problems emerge as a priority. You waste your effort on something that looks useful early on, but cannot eventually be applied. Needless to say, the change is permanent and there is no point complaining about it. But it makes your team reflective; mournful for great ideas that are gone forever.

What would a more-than-slightly paranoid tester, who felt these feelings, suggest for a better test prep the next time he is invited?

Consider just a few simple things:

- Plan the test preparation as a sub-project with the customer and try to agree beforehand on how both parties might benefit. (You may not want to invoice the customer for this)
- Make sure there is a team appointed for this cooperation
- Make sure the start date is not too early, not too late.
- Make sure localisation kick-off and testing kick-off are two different milestones

What Is Clean?

As an everyday occurrence, the paranoid or not-so paranoid tester prepares and tunes up the operating system images for testing. It seems obvious what a clean image means, but an experienced tester knows there are a few potential pitfalls.

Voice inside head:

Can I keep the testing OS clean and secure?

Other voice inside head:

Sorry, I cannot reproduce your bug. Is there something wrong with my clean system?

There are ways to be sure your work is protected and your test machine is not littered with installations from the company logon script, which installs many 'useful' things. Make sure no antivirus gets installed on the machines, or other internal software. Verify with your admin team as to what is the best topology to keep the testing domains secure from outside attacks. You can use a test domain, which is protected from the outside world and has a trust with the production domains. Allow access only to key sharepoints in the production domain, which are indispensable for the team. The test domain should work with your test servers only and admin rights are usually granted for the testers who work with them.

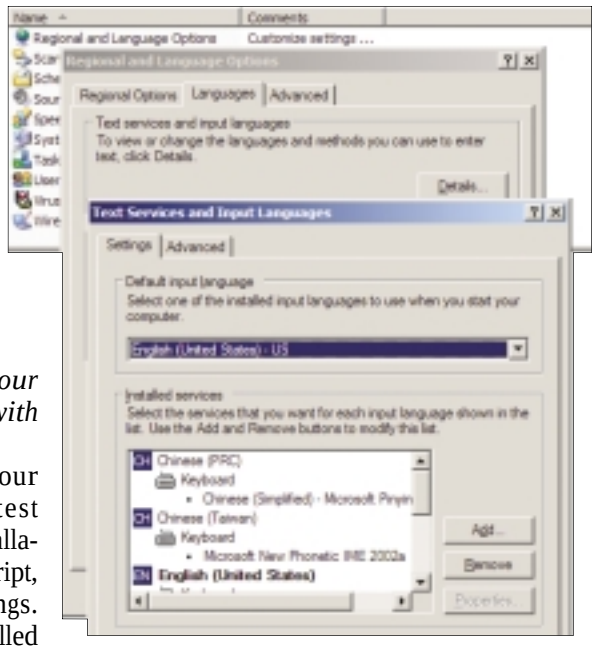


Fig. 1: US Windows XP environment with Asian language pack and inputs installed

Voice inside head … again:

Should I allow others to edit the image?

Other (pesky) voice … again:

If you use a disk image casting server in your company to store and manage system images, it pays to have the important project images set as "read only" and allow for documented copies. It reduces the risk of your perfect image to be undesirably modified or even corrupted.

Voice of doubt:

How do I know it is the right language?

Voice of reason:

Make sure you are testing on the right language. You have tested the product on 3 languages already and now you run on the fourth, say one of the 10 new EU entrants — Slovenian — using an image someone else created. If you test the language for the first time and you are not a proficient speaker, make sure it is Slovenian by learning (at a minimum) basic phrases of the language you are testing: Yes, No, OK, Cancel, Next, File, Edit.

Have the specific keyboard layout in front of you, as well as the national language standards (see www.microsoft.com/globaldev/reference/keyboards.aspx) or use a utility that displays the graphic keyboard layout.

So, what are the things that a good clean system should NOT contain? Install nothing but the product on the clean system, unless there are other requirements. It is wise to have just the localised input language installed on the machine and not any additional reference keyboards. Fig. 1 shows US Windows XP environment with Asian language pack and inputs installed. Though the Simplified Chinese input is not turned on, the files are there on your machine and can have impact on the product you test, not to speak of accidental keyboard switchings.

Tip: Pretend you only speak or write the language you are testing.

Coverage and Team on a Global Project

Say you have a web service product in 11 key European and Asian languages and relatively high-level test cases. The customer puts their trust in you as a global, experienced testing partner and would leave the language-specific approach up to you. Often, they do not make a distinction between French and Traditional Chinese testing and they expect the same level of quality and efficiency on each language. (Perhaps, though, double-byte support and Asian internationalisation has already been identified as a potential risk). As a senior tester or test project manager your job is to decide on the right approach in the test plan. Do Chinese Traditional, Chinese Simplified, Japanese and Korean (CCJK) require Asian language speakers to do the testing? Do you need to split the team and test in several parts of the world?

Questions to consider:

- Is the main task to verify localised product user interface and functionality with arbitrary localised input in the local environment?
- Does your analysis tell you the largest amount of problems would be User Interface (UI) problems?
- Could a large part of testing be effectively covered and supported by suitable tools — UI checkers lke Tool Proof? Automated scripts?

In this case there may be little purpose for a Korean speaker testing Korean. The value here is that a smaller team can do

the whole job and leverage more languages, be it in Asia or Europe, without impairing the product quality.

Are text input methods or dialects important during the testing or are there features like "search" to be tested?

Is any linguistic testing in play, i.e. do you need substantial contextual verification or identification of specific localised text in combined or dynamic features of the product? Does the tester have space for variation and decisions?

The point here is that if the test cases do not read like an idiot box (just do what you are told), you should rely more on tester experience and a proficient speaker testing Chinese could be an advantage. Inasmuch as they can uncover serious problems not directly described by test cases, they can decide to put a low priority or ignore verifications that do not apply for their language (for example, alphabet order or specific formatting). A good solution here could be two international teams; one in Asia, one in Europe, integrating the geographic area language experience. If your process and project management is well synchronised, distance is not a real problem: you can appropriately answer customer needs and maintain value by having the right people do the right level of testing.

Roman Civin, a senior manager with Moravia IT (www.moravia-it.com), oversees the testing division in Moravia's headquarters in the Czech Republic. He is responsible for European and Asian functional and linguistic testing services, including localisation QA, automation and internationalisation. He can be reached at RomanC@moravia-it.com

Managing Translation for Communities in NYC

Debbie Folaron & Igor Vesler

Reading through a form or application issued by a government or similarly large bureaucratic organization is a perplexing task. The typical scenario involves reading once, twice, three times, only to scratch one's head in disbelief and wonder what indeed is being requested in the endless maze of little boxes and spaces squeezed in between the fine print.

In the myriad forms, accompanying explanations and notices distributed to various communities needing social services throughout the United States, we find a slew of impressive acronyms and definitions meant to clarify them merged with abundant doses of legalese that can often defy the logic of non-bureaucratic life in the everyday world. What problems arise when the seekers and recipients of social service aid are the diverse ethnic immigrant communities residing from one coast to the other? How do they understand the concepts, institutions and categories being referred to in the application forms, notices, brochures and other pieces of communication that reach them?

Questions like these emerge acutely for translators and translation companies serving the translation and documentation needs of major city, state and federal institutions and organizations. These large organizations rely significantly on complex administrative and bureaucratic procedures in order to respond to the multiple demands and requirements made on them by the legislative branch. They are beginning to cater increasingly to the wide range of ethnic immigrant communities constantly repopulating and reconfiguring the demographic landscape of the country. In a cursory look at the official Web site for the national Social Security Administration we find relevant information translated into Arabic, Armenian, Chinese, Farsi, French, Greek, Haitian Creole, Italian, Korean, Polish, Portuguese, Russian, Spanish, Tagalog and Vietnamese. How might a single city — often with far fewer means than the federal government — be coping with this? In the New York City area, for example, which includes Manhattan and the surrounding boroughs of Brooklyn, Queens, Staten Island and the Bronx, the population is in fact more racially, ethnically and culturally diverse than the United States as a whole. How do local translation companies deal with the challenges these documents and city agencies pose?

History From the Big Apple

When did translation services become an important legal issue for city institutions based in New York? The New York Legal Assistance Group (NYLAG), the Puerto Rican Legal Defense and Education Fund and Make the Road by Walking (MRBW) made New York City history in 1999 when, representing the various immigrant communities, they filed a class action lawsuit with the federal Office for Civil Rights against New York City. The lawsuit accused the city of unlawfully discriminating against non-English speakers by failing to provide interpreters and translated documents. As a consequence, some organizations were mandated to ensure equal access to social service benefits to the LESA (limited English-speaking ability) community. The community applicants and recipients were to be furnished with information, notices and application materials in their respective native languages.

What languages are these? In 2002, representatives of the Mayor's Office of Immigrant Affairs and Language Services (MOIALS) were guest speakers at one of the monthly New York Circle of Translators (NYCT) meetings, and they provided some interesting statistics. Out of a population in New

York City of over 8,000,000 people, 36% are foreign born and represent more than 200 ethnic groups. Most immigrants reside in Brooklyn (35%) and Queens (31%). No one single ethnic group makes up a majority in New York City. Citing year 2000 statistics from the Board of Education, Office of Bilingual Education, the MOIALS representatives stated that the languages most widely spoken were Spanish, Chinese, Russian, Haitian Creole, Bengali, Urdu, Arabic, Korean, Punjabi and Polish. The languages cited as "emerging" were Gujarati, Hindi, Farsi, Turkish and Central Asian languages, Albanian, Wolof and other West African languages, Vietnamese and Cambodian.

Multilingual Needs for Social Services

One of the city organizations that began to implement procedures for translating its own documents was the Human Resources Administration. Documents include application forms, notices relating to status of benefits and services, notices relating to customer rights, changes in regulation and language assistance posters. In a press release the city agency notified the New York City community that in order "to make access to Food Stamps easier for immigrants and refugees who have problems reading English," it had "translated more than 60 of its own Food Stamp forms and letters into eight additional languages" (other than English and Spanish): Arabic, Chinese, French, Haitian Creole, Korean, Russian, Vietnamese and Yiddish. Free interpreter services could likewise be arranged. As such, it had responded to the lawsuit against the city by beginning to comply with the provision of guaranteed rights and services according to Title VI of the federal Civil Rights Act of 1964. This Act states that "no person in the United States shall on the ground of race, color or national origin, be excluded from participation in, be denied the benefits of, or be subjected to discrimination under any program of activity receiving federal assistance."

From Paper to Practice

What looks easy on paper is not necessarily so in actual practice — both conceptually and technically.

Multilingual posters are now visible in various social service centers throughout the city, as one of my visits recently testified. These "simple" posters notify readers in their native languages and scripts that interpreter and translation assistance will be provided for them should they need these services. On one particular poster, the short paragraph is written in approximately 20 languages and fits on the front and back of one sheet of paper.

For those of us who have had experience working with non-Western languages, it is easy to see that the challenging part is to get the myriad scripts and fonts to cohabit the same page in one word-processed document defined by precise formatting requirements. The Arabic, Hebrew and Yiddish are bidirectional. The Chinese and Korean are double-byte. The Slavic languages are in Cyrillic or need additional diacritical marks for those languages using extended Latin character sets. Languages such as Khmer, Lao, Urdu and Vietnamese require compatible fonts, which sometimes materialize only after considerable energy has been expended to hunt them down. In addition, English acronyms and telephone numbers are retained, and this complicates word processing for some languages on one and the same page.

Finally, one can imagine how producing a document in one word-processed file on the computers in one office does not guarantee that it will be called up in exactly the same way — on screen and paper — in other offices be it as a Word or PDF file, precisely because of the special characteristics of fonts needed for the more "exotic" languages.

A Babel of Languages and Contexts

As our experience in the field has borne out, just as the technical issues are fascinating and challenging for project managers of multilingual documents, so equally are the linguistic ones for translators charged with the task of translating social services documents for the various ethnic communities in New York City. As every translator is well aware, translating into a target language entails envisioning and taking into account the target audience itself. In this case, multiple and decisive factors come into play.

First, the documents were originally designed in most cases for an English-speaking public without a thought given to subsequent translation into other languages. The source language itself assumes a basic understanding of the concepts and institutions proper to the federal, state or city social services system. We will see later how this materializes into real problems when transferring the concepts and terminology into languages other than English.

Second, the documents generally have undergone a series of modifications, updates and revisions, either to conform to newly implemented legislation or to respond to changes within the system. Terminology in the original English versions, which is not necessarily used in the same way by the same person when the documents are being drafted and finalized, is not used in a consistent manner across the board.

Third, the documents are drawn up by lawyers having legal aspects in mind as their highest priority. Legal specialists are in no way different from other professionals in their assumption that others can grasp the concepts and terms of their art as easily as they do. While these documents are designed to inform, notify and inquire, they are crafted and spelled out with the back-door concern of preventing potential conflict with the law. This is in addition to the bureaucratic slang that is typically used and understood internally within any given agency and which poses true challenges for anybody outside the agency to understand.

Fourth, the documents ultimately play a pivotal and decisive role in the real lives of people, most of whom are applying for assistance because they are in desperate straits or incapable of functioning for whatever reason in the marketplace. Real-life decisions are based on the information requested and obtained through the medium of the translated document.

When considering the target audience for the intended translations, the translator is first and foremost presented with

readers who have limited English speaking ability, who are not particularly adept at understanding the legalese of the document language and who are already struggling to grasp basic forms and concepts of a social system that is sometimes radically different from the one left behind. For some ethnic groups, such as Chinese and Russian, the proportion of elderly applicants and recipients of public assistance is higher than for others. This may pose additional problems.

Translating for the Russian Community

What are some of the hurdles to overcome when translating this kind of documentation? Here are a few examples.

(1) The social services/welfare system in the United States is designated by an elaborate schema of categories and sub-categories that divide people ("individuals") into groups and classify them. Most of these classifications rely simultaneously on a combination of two or more implicit concepts, such as family relationships between individuals, their economic relationship (by virtue of sharing the same apartment, for example, or buying food and home supplies) and even sexual relationships (by introducing the concept of partner which is unknown or marginalized in traditional cultures). Therefore, when basic terms such as family and household are used interchangeably in the source language document without explicit definitions, then the word choices in a target language such as Russian or Chinese place a translator between a rock of being truthful to the original and a hard place of properly rendering a culturally alien concept to the target audience. One might choose terms [семья, семейство, родственники] to indicate the kinship-relation of family or [совместно проживающие, ведущие совместное хозяйство] to describe the household as individuals living together, while sharing and contributing their earnings. Since "household" refers to a concept that is less than natural to Russian or Chinese native immigrants whose social cell is normally a kinship-based family (whether immediate or extended), it can provoke confusion. Problems ensue because the source document or form design simply leaves no space for any comment, description or explanation with regard to the term used.

(2) The legal system in the United States and its expression in terms of the social services infrastructure likewise imply an understanding and acceptance of the physical person or individual, rather than the collective body, as the "center." It refers to the individual based on certain attributes and within the context of an individual's relation to the community and state. Relationships of kinship are a case in point. For example, a number of forms use such terms as *custodian*, *guardian* or even *legally responsible relative* interchangeably despite the fact that they do not necessarily imply a kinship relationship in the English language of the United States. What to do if respective terms in the target language do contain this relationship? What if they *don't*? Terms that are proper in the target culture might be inappropriate in the legal aspects and, in addition, might evoke suspicion as an inaccurate and unprofessional translation. They will at least provoke confusion and further questions. How will the applicant then categorize or classify himself/herself? By the same token, for terminology that has grown specifically out of the US context — such as *worker*, *case worker*, *participant* or *recipient* — careful selection of native words or descriptive versions of them must be judiciously employed in the target language text.

(3) The documentation that a government agency requests as proof or evidence of status may also be totally irrelevant or unknown to the target language group. Examples include questionnaires for recent immigrants that refer to service in the US Armed Services or railroad retirement, or addressing Jewish synagogues, Muslim mosques and other non-Christian religious establishments and their officials as *churches and missionaries*. One typically thorny concept is the notion of a *funeral agreement* (or a letter from a *funeral director*), which is difficult to render in Russian, for example, when funds for funeral services are generally entrusted to a local church, religious establishment or clergyman when no legal heir is available. Or how the terms *death benefits* and *survival benefits* (again, used interchangeably) should be rendered for audiences whose native social welfare systems simply lack such concepts?

Keeping the End User in Mind

It is true that in general the documents are designed in the source (English) language to be "all-encompassing," even "universal" in a sense, in order to cover the whole spectrum of cases possible. This could be considered a most laudable aim, but it is also precisely this goal that tends to "translate" into authentic confusion or embarrassment for the end-users when translators are asked to render the text into target language-cultures that deal with certain issues in very diverse ways. Health-care service documents are a case in point.

Let's imagine Masha, a recent immigrant from Russia, as she encounters for the first time a health-related brochure distributed in a local hospital. The text, in usual straightforward English-language fashion, informs her of how one should deal with certain maladies or emergencies:

You have lead poisoning.

You should be hospitalized immediately.

Rather than consider the text as purely informational, helpful and informative, she may very well wish in her heart that she had never read the line. Phrases expressed so affirmatively and addressed in such a direct tone traditionally carry much weight in her native language. Words, once pronounced and written, have an almost magical power conferred on them. Now she might be cursed with the affliction she has just read about!

Another example might be Malika, a recent immigrant from Eritrea. She has just been given information on pre- and post-natal care in a local health care clinic. As she begins to leaf through the translated brochures, she is shocked at the exceedingly explicit text describing and showing female physiology as one might find in a textbook for medical students. She may even close the brochure, never to open it again, especially in the presence of her family. This kind of layout is considered by some cultures to be extremely offensive, bordering on the obscene, not the least because it is printed and mass-produced for such overtly public distribution.

And yet another example could be Nguyen Thi Minh, who has recently arrived from Vietnam. At her latest visit to one of the local charitable organizations, she was given a brochure on proper nutrition and nutritional needs to assist her in learning how to eat "correctly." The brochure was designed and crafted in English for an American audience relying on the services of such organizations. It contains basic assumptions (like that of spending too much time in fast-food restaurants) and implications that, in fact, run counter to the healthy eating habits of Vietnamese traditional culture. The intended message of the source-language culture translates into quite the opposite message when rendered into Vietnamese or for other groups such as Indian, Russian and Jewish.

Are There Any Solutions?

What ultimately can be done, then, to facilitate the translation and "localization" of these types of administrative documents from the English source language into the various target languages? The ideal solution, of course, is to radically overhaul the source documentation system and procedures from within so that it operates on a totally different set of precepts and standards, thereby facilitating subsequent translation into languages other than English.

Looking at the scenario through rose-colored glasses, we might see direct participation in the initial source language document creation stage by language specialists and documentation consultants who have direct exposure every day to both the communities being targeted and their language preferences. However, looking at the crystal ball instead, we can safely predict that most institutions and organizations will have neither the means nor the will to carry out this kind of massive overhaul.

What "partial" solutions and immediate patches can there be, then, for translators and managers of these projects in anticipation of more radical steps by these organizations and institutions?

First, *create and maintain a multilingual terminology database* from which glossaries can be easily extracted for each of the languages and from which a comprehensive thesaurus can be compiled to establish the relationships between terms (to categorize and synonymize them at least). This would ensure (1) consistency and uniformity in understanding and rendering English terms in the target languages and (2) cross-language consistency and uniformity of the translated documents. Likewise, examples of context-heavy English terminology should be included in the database with thorough explanatory notes and definitions of their respective renderings into the target languages.

Second, *employ translators and editors who thoroughly understand the system, context and language of the source language documents* as well as the realities of social benefits and social welfare systems in their native countries and, at the same time, are exposed to the language of their ethnic community in the United States. They should use proper reference materials and documentation found in their native countries, when at all possible, as they are rendering the document into the target language. This will ensure accuracy and the proper localization of each document for its intended target audience: the local ethnic community.

Third, *collaborate as closely as possible with administration project managers and reviewers* by soliciting their input and feedback and by relaying it to the translators. Attempt to influence the document creation and revision process by submitting comments on document contents; implementing version control mechanisms; advising the client of peculiar, inappropriate and irrelevant concepts, inconsistent terminology, quasi-repetitive portions of text; and so on.

Fourth, *facilitate permanent client education* by making it an articulated fundamental goal for all documentation experts and professional translators, as well as for the translation company's project managers.

The appropriate transfer of both language and cultural context for documents like those pertaining to social services, community medical services and educational organizations is complex and laden with potential landmines in terms of authentic comprehension and communication. A "simple" mandate to the translator to "translate at a fifth grade reading level" those documents with complicated legalese or an assumed understanding of target social infrastructure and structures remains enormously challenging. A "simple" mandate to the project manager to have all text, regardless of language, conform to a specific font size and formatting proper to the English language — on a page that clearly has not one iota more of space available for the source language — may provoke enormous complications and waste valuable time just to perform the technical acrobatics necessary to complete the project.

Truly, a more defined role with active participation and visible representation by professional translators, linguistic/cultural consultants and expert project managers is a necessary first step to combating these problems and overcoming the hurdles so that a more effective and beneficial solution can be found for all: the government institution itself; the recipients of the target language documents; and the companies and translators contracted to bring these complex multilingual projects to fruition.

Who knows? All of us native English language speakers might even derive some benefit by being able to read clearer and more consistently uniform source language documents and have to pull fewer hairs during a process that is typically cumbersome, time-consuming and sometimes downright frustrating. ■

Debbie Folaron is manager of language and technology at Eriksen Translations. She can be reached at debbie @eriksenine.com

Igor Vesler is a freelance Russian and Ukrainian legal and technical translator and interpreter. He is an independent consultant (Diken Research, New York City) and can be reached at vesler@compuserve.com

New Organizations Serve the Localization Industry

Nancy A. Locke

A relatively new industry, localization is nonetheless showing signs of maturity. The emergence of new organizations comprised of distinct constituencies is one such sign. Efforts to define, align and then realign those constituencies to improve representation — notably the merger now taking place between the Professional Association of Localization (PAL) and The Institute of Localisation Professionals (TILP) — are another sign.

Since 1990, the Localisation Industry Standards Association (LISA) has represented the corporate members of the globalization, internationalization, localization and translation business communities. The organization has provided the industry with a forum, voice and identity open to service and product vendors, clients, the academic and R&D community and, more recently, individual professionals. As a logical result of growth and the growing diversity and richness of the industry, however, other voices are now making themselves heard.

Professional Association of Localization (PAL)

In an essay published in *MultiLingual Computing & Technology* (#31 Volume 11 Issue 3), Reinhard Schäler wrote, "A strong industry association is necessary for localization to gain long-denied recognition as an industry in its own right." Schäler, director of the Localisation Research Centre at the University of Limerick, Ireland, concluded, "What is needed is for somebody to take on the initiative and lead."

Ricardo Erb, a localization veteran with 15 years' experience, envisioned a new organization of individual professionals whose needs might be different than those of the companies for whom they worked. The focus on individuals also acknowledged the existence and important role of freelance professionals to the industry. Erb and a group of like-minded colleagues founded PAL, a nonprofit organization, in early 2001.

PAL had ambitious goals: to serve and represent those who prepare software and documentation for the world. According to its Web site, PAL's mandate is three-fold: "to support its member translators, localizers, internationalizers, and globalizers, PAL assembles and provides information, organizes professional events, and serves as an advocate for member concerns."

"As the first professional organization for individuals, starting up when and how we did, PAL faced major hurdles," says PAL president Suzanne Topping. "Our virtual team organized and ran an international association without a headquarters, funding, or dedicated personnel. Given these constraints, it was a constant juggling act to bring in new members while developing programs." For PAL, funded solely by membership fees, building membership has been a constant preoccupation. Offering potential members value for the yearly membership fee has been equally important. The success and failure of these two priorities proved to be interdependent.

To date, the association's Web site and a small, informal network of colleagues have been the only member services offered by the organization. Other projects have been too expensive or logistically complex to realize. Growing pains, including a difficult leadership reorganization in early 2002; prolonged economic instability particularly in the high-technology sector; and upheavals in the personal lives of PAL's leadership council made forward movement more difficult.

PAL's existence and, in particular, the organization's emphasis on individual professionals have, however, had an impact on the industry. While a causal relationship is impossible to verify, Topping notes that LISA now offers individual memberships.

"PAL's creation appears to have inspired a variety of people in the industry to take action," Topping says. "It helped the globalization community realize the need for cooperation and for support organizations at various levels. This change is an indication of the maturation of the industry and recognizes the variation of needs by different groups."

PAL's mission and commitment to localization professionals remain strong; however, at the beginning of 2003 the association began evaluating more effective means of realizing them — specifically, the possibility of merging with another professional organization that shared its fundamental mission. The Institute of Localisation Professionals (TILP) seemed like a logical partner.

The Institute of Localisation Professionals (TILP)

TILP is the most recent initiative of the Localisation Research Centre (LRC) in Ireland, which has been the nexus of research and educational support for the localization industry since 1995. Formerly located at University College Dublin, the LRC is now hosted by the University of Limerick. With the support of key industry players, the organization has been the source of many industry-related initiatives. LRC has developed a diverse offering of industry-related services including the "research and development and evaluation of localization tools, the establishment of a Localisation Tools Library, consultancy services, education and training and the publication of a regular newsletter." LRC is also home to the European Localisation Exchange CenTre (ELECT), a two-year project funded by the European Union's eContent Programme.

In March 2002, Schäler invited a group of professionals to a meeting at Dublin's Merrion Hotel. On July 30, 2002, TILP was formally registered as a nonprofit organization created "to develop professional practices in localization globally." While TILP has recruited corporate sponsors such as Alchemy Software Development, Bowne Global Solutions, Lionbridge, Microsoft, Novell, Oracle, Symantec and Veritest, the needs of individual localization professionals form the core of TILP's objectives.

"The LRC is one of the research centers at the University of Limerick. It is owned by the university," Schäler explained. "TILP was established as the professional body of the localization professionals — similar to other professional bodies (bar council for lawyers, similar bodies for engineers, architects, doctors)."

TILP offers three types of membership: Associate, Professional and Fellow. Anyone with an interest in localization may join TILP as an associate member. Professional memberships, however, are subject to review and recommendation by the Professional Membership Review Committees and approval by the TILP Council. Within the professional category, TILP identifies three distinct groups: language, project management and technical professionals. Prospective members must complete an application and the appropriate supplementary information form, which requires providing background on education and professional experience. The supplementary information must be verified and signed by two professionals who must themselves meet certain criteria. The TILP Council also names Fellows — for example, experts in their fields who will act in an advisory capacity.

The TILP membership structure embodies two of the organization's most important goals: professional development and professional recognition. A related goal, the development of a "certified localization professional" (CLP) standard, is also a top priority. TILP is currently planning a series of meetings in both Dublin and the United States.

TILP and PAL to Join Forces

In early 2003, the PAL leadership council approached TILP with the idea of a merger. Both organizations quickly found common ground. The two leadership councils felt that, given the proliferation of industry organizations, operating as two independent organizations was no longer an efficient way of reaching the individual professionals who were counting on PAL and TILP to bring about positive changes in the industry. The two organizations would be more effective as a single entity in meeting the needs of a shared constituency.

In mid-February, the two associations agreed in principle to join forces. The newly merged association will operate under the name TILP. Details of the merger, including membership transfers, are still being worked out.

There are many reasons for the merger. For one, both organizations address the concerns of individual professionals rather than corporate service vendors or clients. While the interests of both localization professionals and the companies for whom they work both on the client and the provider side necessarily intersect, other areas of concern are unique and of high importance particular to each group. In this sense, PAL and TILP are more akin to traditional professional organizations and guilds such as translators' associations than to LISA or GALA.

The merger of PAL and TILP will increase the size and geographic reach of the resulting association, eliminate organizational redundancies and create a synergy in terms of leadership and professional expertise.

Globalization and Localization Association (GALA)

Unlike TILP or PAL, the Globalization and Localization Association (GALA) represents companies in the internationalization,

localization and globalization industry. GALA's mission is "working together to share information, fostering innovative ways to promote ourselves and our industry, and offering our clients unique, collaborative value."

In an e-mail exchange, Hans Fenstermacher, president and founder of ArchiText, Inc., and a founding member of GALA, addressed the emergence of new industry organizations. In particular, he spoke to the notion that a "partitioning" of the industry was under way. Favoring the term *segmentation* to describe what is happening, he wrote, "We have seen far too much fence-building in the history of our industry, which, frankly, is a sign of our industry's immaturity. I can't speak for other new organizations, but GALA's aim is to represent not the entire industry (in my view, that's logically impossible anyway), but the vendor community, those that provide the products and services. We believe that vendors have interests, problems and objectives that differ from those of other constituencies in the industry, such as clients." While Fenstermacher thinks that "we are potentially creating too many organizations to serve needs that are too small," he added, "Time will tell, but we certainly think there is room in our industry for several organizations."

Close member cooperation necessitated the development of a code of conduct and a mutual non-disclosure agreement. The membership packet available at the organization's Web site includes these documents. Membership in GALA is open to any "bona fide company providing translation, localization, internationalization, or globalization products or services, including tools developers, training suppliers and consultancies." Fenstermacher explained the *bona fide* criterion as "meant to distinguish GALA from the mission of other organizations that focus on individuals or loose affiliations of individuals. Everything about our organization is focused on the interests and objectives of companies in our industry."

In April 2003, GALA will celebrate its first year anniversary. By any measure, the organization has been a resounding success. The organization launched with 15 companies committed to its mission. To date, GALA has nearly quadrupled in size and now boasts 56 members from more than 12 countries and four continents. In February 2003, GALA accepted the Industry Association of the Year award sponsored by ClientSideNews (www .clientsidenews.com).

"GALA is making plans for several smaller regional events," Fenstermacher says, "ways to join other conferences and events or opportunities to co-brand events with other organizations. We are acutely aware of our members' concerns about return on investment for travel and conference fees, and we are following a best-value approach to meet those concerns. One of the events later this year will also be a general members' meeting."

The Association of Language Companies (ALC)

One of the most recent organizations to emerge, the Association of Language Companies (ALC) probably has been the longest in development. The impetus for the formation of the ALC came from various members of the Translation Company Division of the American Translators Association (ATA), who felt that the development of meaningful policies relating to quality standards and other issues of interest to translation companies might best be achieved by an association independent of the ATA.

Formally founded on July 31, 2002, the ALC still struggles with the development of a quality standard or norm. For over three years, as individuals, ATA members and now as ALC members, the founding members have consulted standards organizations in Europe and Canada, including the Association of Standard Testing of Materials (ASTM); Association of Translation Companies (UK); British Standards Institution; Comité Européen de Normalisation (CEN); Deutsches Institut für Normalung (DIN 2345); European Union of Associations of Translation Companies (EUATC); and the International Organization for Standardization (ISO 9000/9001).

The ALC has assumed a more active role in the international standards dialog. In December, ALC was the only US organization admitted as an observer to a series of conferences sponsored by the CEN to be held in various European cities. The goal of this work is simple: to define a more specific quality control system for professional language service providers. In addition, the ALC intends to serve as a focal point for developing and maintaining professional guidelines; maintain and share information for the growth and benefit of members and their clients; educate and inform interested parties; and provide forums for discussion and learning.

According to Suzanne Robinson, one of the ALC's founders, "member adherence to the future ALC norms will be defined by a strictly voluntary professional oath. The norms will provide a guideline for companies to affirm their serious concern for quality, from qualification of contracted language professionals to conscientious project management."

ALC offers three levels of membership. Active members are "companies actively involved in the sales, marketing and production of language services." Active members enjoy all ALC benefits as well as voting rights (one vote per company). Affiliate members include companies without an active US presence, independent organizations and "educational institutions and other professional organizations involved in the translation/interpretation field." Affiliate

The New Organizations at a Glance

Organization	Web Site	Membership Classes & Dues
Professional Association for Localization (PAL) *merged with TILP in 2003*	www.pal10n.net	Individual — $75/year Student — $55/year
The Institute of Localisation Professionals (TILP)	www.tilponline.org	Associate — €50/year Professional and Fellow — €100/year
Globalization and Localization Association (GALA)	www.gala-global.org	Company — $350/year
Association of Language Companies (ALC)	www.alcus.org	Active ($450/year), Affiliate ($275/year), Vendor ($2,000/year)

members have restricted benefits and do not have voting privileges. Vendors to language service companies that wish to support the association's efforts to promote quality standards in the United States may join, and their logos will appear in the association newsletter. Vendors do not have voting privileges, but they have access to certain benefits.

The ALC's first annual conference, to be held in Portland, Oregon, on June 12-14, 2003, will address a number of topics related to the business of providing language services as well as standards and quality issues, technology and sales. ■

Nancy A. Locke is a freelance writer and translator and chairman of the PAL board of directors. She can be reached at nancy_locke@sympatico.ca

GAMES LOCALIZATION: PRODUCTION AND TESTING

Helen Trainor

Computer games have become a worldwide phenomenon, and successful gaming companies are finding that localization is now a mission-critical aspect of their business models. When the industry started out with Space Wars and Pong in the 1970s, localization was thought of as a support or marketing issue rather than a development issue if it was considered at all. Localization was usually limited to translating product packaging, documentation and marketing collateral. There just wasn't a need for software localization when the game was essentially moving dots on a screen.

But much has changed in the gaming industry, and today's world looks very different to developers. Gaming has developed into a global industry that is on the cutting edge of 3-D graphics and coding technology. Many of the hardware and graphics advances in the consumer technology industries have been driven by the increasing number of polygons and speed of rendering demanded by gamers.

Games developed for Asia, such as Studio ego's Angel's Feather *(localized by Symbio), often have a fantasy look and feel*

Along with the technical advances in gaming, the economic stakes have gotten much higher as well. Many titles now require blockbuster sales figures in order to justify their Hollywood-style development budgets. In order to achieve these, gaming companies are increasingly looking to expand into international markets.

The global demand for computer games has also exploded. Console games currently dominate the industry, but demand for other formats is increasing. For example, a recent IDC Internet user survey ("IDC Sees a Big Boom in Online Gaming," January 16, 2003) shows that 50% of Internet users in Korea are on-line gamers, as well as 43% in China, 39% in Singapore and 35% in Hong Kong. According to IDC's research, as more end users in Asia/Pacific come on-line, the number of Internet users who regularly play on-line games will continue to increase.

To tackle the increased technical complexity and to ensure that titles will be able to generate those much-needed international revenues, gaming companies are now planning in the early stages of game development for international releases. Although you might initially assume that game localization is not much more complicated than the standard software project — especially action and arcade-style games that don't contain much dialogue or text as opposed to text-heavy role-playing games (RPG) — the process of localizing games for multiple countries introduces some interesting new challenges.

At the most fundamental level, games tell stories. The localizer's challenge is to make these stories resonate for different cultures.

Internationalizing the Underlying Technology

To begin with, as in all localization projects, planning is a critical success factor. The first order of business is to ensure that the game is developed so that it can be easily localized down the road. Unfortunately, in many cases, games require extensive "rip-and-replace" strategies during the localization phases because fundamental issues were not addressed from the very start of development.

The first major technology issue that needs to be addressed during product development is enabling the game for double-byte locales. When localizing English or other single-byte language games for the Asian languages of Chinese, Japanese and Korean (CJK), you must enable the game to input, display and process double-byte fonts.

Games developed for single-byte languages often use graphics to replace the system font. This adds to the aesthetic appeal and works well because there are a finite number of characters in Western alphabets. However, when localizing into Asian languages, the underlying text display routines may need to be redesigned because Asian languages are character-based. As a practical matter, it is almost always cost prohibitive to create a graphic font for Asian languages as there are thousands of characters. An easy way to handle this problem is to use the built-in double-byte fonts bundled in Asian Windows. Windows has APIs that emulate and create the font so that the double-byte font can be used through other common APIs such as TextOut.

Platform porting can also be difficult. During development, it should be decided if the game will be ported to other platforms and, if so, which ones. Porting a game from a PC format (running a Microsoft operating system) to the X-Box is relatively easy since the X-Box is developed on top of similar Microsoft operating system technology. However, there are more challenges when porting a PC game to a Sony PS2 or Apple Macintosh environment. Although tools are available that claim to port code to and from various environments (and more are being developed every day), it is important to allow time to perform significant levels of testing and recoding after porting, especially when dealing with complex codes such as Direct X.

Another technology issue that occurs in PC and on-line gaming titles involves input methods, as they need to accommodate the different keyboard layouts and input devices available in various locales. When it comes to input method editors (IMEs), we recommend supporting just the most popular IME in the target market and perhaps one or two other popular IMEs to maximize the user coverage. Some browsers, such as Microsoft Internet Explorer; development tools, such as Visual Basic; C++ classes, such as MFC; and Java classes, such as Swing, may take care of this issue natively. Many English applications, however, especially some written in C/C++, are not IME-aware. For Asian markets, Asian Windows has built-in IME support; however, the application should talk to the built-in IME through Windows IME API (for example, ImmIsUIMessage) so that IME support can be enabled. Lastly, most console platforms don't come with a keyboard, so a well-designed IME should also consider the interface usability, such as a virtual keyboard.

When it comes to the game's story line and characterizations, there really aren't any technical solutions to speak of. But whenever practical and where it is relevant to the spirit of the game, developers can use universally understood icons. In some offerings, such as simple arcade titles, icons might eventually replace the use of language to communicate across cultures.

Culturalizing the Game's Story and Characters

Culturalizing games can be one of the biggest challenges of all, particularly in RPG and plot-driven games. Games are created for entertainment value and therefore tend to be carefully crafted to suit a particular audience. This can be seen most pointedly in game characters. Western game characters, for example, are usually more "adult-like" and have more pronounced physical characteristics (think of He-Man or Lara Croft), while Asian characters typically emphasize more child-like characteristics (think of Japanese *anime* or *manga*) and have more of a fantasy look and feel. For certain titles, localizers may find they need to recreate characters for the target market. This can be a very involved effort. Many artists and character designers are very strict in their control of the use and interpretation of their designs, and typically the designs cannot be altered at the localization stage without the consent and often cooperation of the original designer(s). Unlike most business software, there is often a certain amount of artistic pride involved, and artists can be extremely protective of their work. As such, as much as the technical component, the logistical component of getting a sign-off from all related parties is an issue which has to be addressed at the beginning of any game localization project. This has become more of an issue in recent years, as games have increased their artistic content, and from this perspective is similar to the way movie rights are handled.

Additionally, different countries have different tolerances — and sometimes laws — for the amount of violence and sex in games. It is smart to be aware of the ratings systems in place in various markets, such as the ones endorsed by the European Leisure Software Publishers Association (ELSPA) and Entertainment Software Ratings Board (ESRB). Some countries have outright bans on graphic sex and violence.

Germany in particular has strict policies about the amount of graphic violence that can be shown in games. There are ways of getting around this, of course. Since red blood is considered gore but green "blood" is not, changing the color during the localization process may make all the difference. In some cases, adding gore is the issue. Specifically, US-developed games are much more liberal in the amount of gore used than many successful Japanese games. When the Japanese games are localized for US release, the amount of

gore is often increased to meet audience expectations. Building variable gore settings into the game at the outset is a key best practice. For example, in the Far Eastern Edition of *Baldur's Gate* by Bioware (localized by Symbio), the local requirement is that the default be set to "no gore," but the player then has the option of adjusting this setting. In this case, Symbio met the local restriction without changing the player's options.

This brings up a point about differences in default settings. Different markets have different preferences for default settings. For example, in driving titles, US and European markets desire different vehicle handling settings in driving titles. For games that increase the difficulty as the player achieves different levels, the initial difficulty levels are sometimes lowered for the Japanese market as the game audience tends to be younger.

The differences in the laws and cultural sensitivities with regard to sex and nudity and issues such as the implication of child pornography legislation must also be understood and addressed. For example, depictions of genitalia (real or animated) are expressly forbidden in Japan, and such images must be "doctored" for distribution in Japan. On the other hand, Japanese adult *anime* games often have scenes of sex and nudity involving animated/computer graphic depictions of characters who from facial features, attire and story setting could or would be interpreted as being below the legal age of consent. While this type of depiction is perfectly legal in Japan (as long as the games are distributed to adults), many other jurisdictions take a different view. Exporting such games requires, at minimum, alterations to the explicit scenes. Often the plots are completely rewritten without the adult component, in effect just reusing the character designs and other artistic material developed for the original game (together with the underlying RPG engine) in the distributed game.

Translation and Voiceovers: The Fun Part

Simple text translation in games can be challenging, as it can be argued that gamers have their own language. It is important to use native speakers who are familiar with the slang and game terminology within the target country. Elements such as quizzes, jokes, puns and even story lines may have to be redesigned rather than just translated. For online and RPG titles, many of which are of a fantasy/sci-fi genre, genre-specific terms are difficult to translate. For example, how would you translate Orc into Korean? Furthermore, if animations or graphics are associated with the text, text expansion can become a real problem.

Voiceover localization is a tricky undertaking if not well thought out prior to the beginning of the project. Depending on the genre of the game or the market requirement, you may even want to skip this step. For example, you may choose to leave the voice unlocalized for a strategy game or simply to add subtitles for movies included in a game. On the other hand, you can give players the option of choosing to hear the English or base language and see localized subtitles or to hear localized voice and see localized subtitles. Using multilanguage options is often not done, however, especially within the games console market, as this type of feature can encourage cross-border black/gray market sales and may weaken the control of the game vendor over the distribution and possibly decrease the profitability of the localized game.

To localize voiceovers, you must begin by finding the appropriate voice talent. The importance of having the right translation team for the job cannot be overemphasized. As with any business application vertical, some translators specialize in games. It is important to find actors who will take into account the target locale and its culture. If phrases will be used repeatedly in response to different situations, having the actor maintain consistency in intensity is crucial. And don't forget that background noise needs to be consistent with the action on the screen.

Timing is another issue with voiceovers that often affects the movie industry as well as the gaming industry. Getting the voiceovers to match the screen if the end user is seeing a character speak is not a simple undertaking. If you are localizing an English product into German, the German speech will normally take 30% longer. You will have to either speed up the voiceover or lengthen the corresponding screen time. A further complication is that in many Asian locales you will have the opposite problem, as the speech time will tend to be shorter than English and other western languages. To fix this, you can have the speaking character animation/movie run until the text file has ended, or you can have the program read the file's length and match the timing. Otherwise, players will notice this as a glitch.

When using voiceovers, it is important to have a detailed matrix of each voiceover, the actor, the base text, the translated text, where and how often it appears in the game and any on-screen actions or issues associated with that recording. Fixing problems with voiceovers late in the localization cycle can be difficult because you will need to use the same people for consistency, and they may not be readily available. For this reason and because you want to ensure that testers have found most of the text errors already, it is smart to leave the recording of voiceovers until late in the localization schedule.

Testing Localized Games

There is no one testing plan that will work for each game since the locale and platform of each game are unique. Like other types of software, games need to be tested for compatibility, functionality, performance and translation verification. However, depending on their locale and platform, games may also require testing for additional items. These can include graphics (important for 3-D games), sound card/engine and special effects (does the building blow up when the bomb is detonated?). Story-style games also may require more testing to ensure the consistency of the changing landscape of the game (does the world around the character move and adapt as it should?) and that the roadmap of the game works properly (when the player is given three choices of action, does the game map the appropriate choice to the appropriate response?). A solid and detailed testing plan and an experienced team are essential to success. Be sure to include not

only the types of things you want tested, but take into account the many platforms and environments that will be used to run the game.

Many firms make the mistake of testing for the minimum qualifications to load and play the game and then assume that the game will work with more sophisticated peripherals and systems as well. It is important to use many different types of machines (both Tier-1 and "homemade" configurations) for PC games and many connection speeds and methods (ISPs/Modems/Cable/DSL/T1) for on-line and interactive console or RPG games. The constant evolution of gaming and computer accessories necessitates a more focused testing plan. Another issue that needs to be considered, especially for PC games, which can be more difficult to test fully due to platform issues (nonstandardized hardware, operating system issues and so on), is the ongoing issue of localized support.

Last, it is important to emphasize that time-to-market is often so important that the number of review cycles is drastically shortened. If the product involves translating a large quantity of text and graphics, quality may be compromised. Keep in mind that Asian markets tend to have higher quality expectations than the United States, so poor or missed translations for these markets can hurt not only a company's reputation, but its wallet as well. Testing early and often is imperative.

Conclusion

By their very nature, games can be more complicated than the average software product. Games contain animation (most offer 3-D), voiceovers, lots of graphics, text and music, all of which may need to be localized across a myriad of platforms and technologies. Story-based games can require localizing stylized things such as landscapes and character appearance as well as everyday things — houses, buses, police cars, telephones and traffic signs.

But the international demand for video games, across all platforms, continues to rise. The success of each game and its ability to launch sequels and become a profitable franchise are increasingly related to its ability to be adapted and appreciated by a variety of cultures around the world. ■

Helen Trainor is a senior manager at the Symbio Group. She can be reached at htrainor@symbio-group.com

Karen Drust, marketing director, also contributed to this article.

The author thanks Shinichiro Fukushige, director of Japanese operations; Ernie Huang, director of professional services; and Felix Chen, senior internationalization engineer, for their assistance. For further information about games localization, readers may contact Jacob Hsu at jhsu@symbio-group.com

ELECT Online Offers Information

Michael Bourke

For professionals working in localisation and internationalisation, the Internet is a valuable resource in terms of documentation. There are a myriad of publications including guidelines, standards and other documents strewn across the World Wide Web. But while there is no doubt that there is a vast amount of information available, actually finding this useful information is another matter. Similarly many companies on the Internet require localisation and internationalisation services but do not know where to look, while localisation companies need to find content producers that require their services.

The first aim of ELECT Online is to provide a 'one stop shop' for finding localisation and internationalisation information, be it guidelines, standards or even a basic introduction to localisation. The second aim is to make it easier for localisation service providers and companies involved in eContent to find each other.

Fig. 1: Elect Awards Index

Launched recently, ELECT Online is an exciting new way for localisation and internationalisation companies to make contact with eContent providers and eContent service providers. It is intended to bridge the gap between the localisation and eContent industries.

ELECT Online is a portal, in that it does not host the content itself but points to where the content can be found. It can be best described as a directory of third party documentation and resources. The portal contains thousands of records describing what a particular document is about, or what a company specialises in, and a link to the particular resource or organisation. Already ELECT Online contains over two thousand detailed records.

Categories

The list of records contained in the portal are separated by category (See Fig. 1).

These are as follows:

- **News & Events.** The latest industry news and events from around the world, including an archive of older news articles and past events. Each news and event record has a brief summary. The event records also include information on the start date, end date, location and the organisation in charge of the event.
- **Software Directory.** A searchable list of localisation tools used in the industry. The list can be filtered to find tools for a particular platform, or that are used in a particular area (i.e. help systems, image management). It also tells the user whether the software tool is available on LOTS , the Localisation Technology Laboratory and Showcase located at the Localisation Research Centre.

■ **Glossaries.** A searchable list of localisation glossaries.

■ **Jobs.** Opportunities offered by various companies in the industry, which can be narrowed down by location, field or keyword.

■ **Learning.** A list of academic, commercial and online courses related to localisation/internationalisation, which can be sorted by location and keyword.

■ **Publications.** A large directory of documents, including books, case studies guidelines and standards.

■ **Research Projects.** These links point to ongoing projects in the industry, as well as completed projects.

■ **Professional Directory.** A huge collection of localisation service providers, eContent service providers and eContent producers. The Professional Directory is one of the centrepieces of ELECT Online (See Fig. 2).

It contains thousands of records divided into three categories: eContent Service Providers, eContent Providers and Localisation Services. Using this page, a localisation company can find a list of eContent producers that might require their services, while eContent companies can browse the list of localisation service providers.

Each section contains a list of all the records currently stored on ELECT Online in that particular category. This list is composed of the title of the record, the subcategory it belongs to and the date the record was added. All this information can be quite overwhelming, so there is a form for filtering the list. If someone wants to display only those translation companies located in Ireland, they can do so instantly. Clicking on the column title of each list (Title, Category and Date Added) will sort the list alphabetically or by date ascending under that heading. Clicking on the title of a record will display the record's full details (See Fig. 3).

REGISTRATION

Visitors can also register themselves with ELECT Online. Doing this will add their details to the LRC's contact database which makes them eligible for LRC mailings and the new LRC electronic newsletter. When a user registers they are asked to choose a username and password. This allows them to log onto the site, which opens a list of options for the user. These include:

■ **Update personal details**. If a person changes companies, phone number or anything else, they can log onto ELECT Online and immediately change their details.

■ **Submit link.** If a user wants to add their company's details to our database, they do not have to mail ELECT Online and ask for inclusion. Once registered with ELECT Online, a user can submit their company information to the portal. All someone has to do is log in, click on 'Submit Link', 'Professional Directory' and fill in the company's details.

In fact, a user can add a link to any of the above sections using this method (See Fig. 4).

In order to ensure that records submitted to ELECT Online are correct, a submitted link does not immediately appear on the site. It must be reviewed and approved by one of the ELECT Online administrators before it will finally be viewable on the portal.

Fig. 2: Professional Directory

Fig. 3: Professional Directory Display

Fig. 4: ELECT Submit Link

Once a record has been approved and added to the site, the user who submitted it can edit the record at any time. So, for instance, if the location of a particular document changes, or there is a more recent version of a book available, the user can log onto ELECT Online, look at the records they have created and update them. This cuts out on delays arising from having an administrator making these changes.

Visitor Submissions

It is this link submission functionality that is the core of ELECT Online. The portal depends on these submissions from its users in order to stay up-to-date. Over the coming months, it is hoped that ELECT Online's traffic will rise to a sufficient level so that the site can become self-sufficient in terms of new information being added.

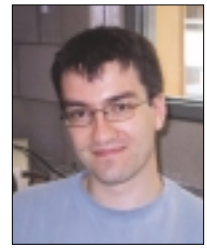
ELECT Online is also intended to create a community not just of localisation /internationalisation professionals, but also between members of the industry and potential customers and academics. To this end, there is a fully featured discussion forum hosted by ELECT Online. It is hoped that frequent interaction between individuals will help strengthen professional ties and encourage business. For security reasons, posting to the forum requires a separate registration to ELECT Online's registration facility.

Internationalisation

ELECT Online is an internationalised site, intended to be an example of internationalisation best practices. Currently the portal is available in English, Czech, French, German and Greek. The links are only available in English at this time but these too will be internationalised in the near future. In order to change the language the site is viewed in, the user simply selects a language from the drop down list box at the top right of each page. The site will appear in the selected language from then on.

The Future

ELECT Online has the potential to be one of the prime sources for localisation and internationalisation information on the web. In order to succeed, it will need to build up a community of users who actively submit links to the portal. In the coming weeks there will be a concentrated effort to push ELECT Online as the primary method of finding localisation and internationalisation information. Once fully established, there are plans for additional features on ELECT Online such as expanded user functionality, advertising opportunities and a weekly newsletter summarising newly added links. ■



Michael Bourke is a member of the Localisation Research Centre and is responsible for the design and development of ELECT Online. He can be reached at michael.j. bourke@ul.ie

A Little Paranoia Can Go A Long Way (II)

Roman Civin

How many bugs are enough? A strange question, indeed. Everyone knows that measuring productivity and quality by bug numbers is, to say the least, a disputable activity, and that productivity in testing is a far trickier variable. Let us open up some of the suitable approaches you can take as a test lead who wants to be clear on bug definitions, avoid unnecessary bugs and help the team learn more.

The assumption is that most customers want only relevant bugs logged. Once you log a bug in the testing phase, everything becomes more expensive, because each bug is managed and meets several people during the cycle. To wit:

- *Find ways to eliminate the bugs prior to product testing:* This can be done through automation pre-testing, proper software QA prior to testing phase, well planned globalisation and internationalisation, and well-timed milestones to name a few.
- *Know which behaviour can be ignored*, so that 99% of what to enter as bugs actually gets fixed. Your "not-a-bug" list could (not) contain these items:

Localised dialogue boxes do not have to match source layout, if they appear OK

Context menu hotkeys

Non-localised components:

Product names

Very trivial cosmetic bugs, e.g.1 pixel misalignments

Functional bugs caused by a missing key component

- *Provide training and feedback to your teams* to avoid duplicate and non-reproduceable entries, not only before the project starts but also as it progresses.
- *Be sure to have the testing types well defined in your testplan*, in order for your engineers to perform the right testing. Otherwise you might end up with dozens of minor language errors that your Turkish tester finds on the product's functional testpass, which you know cannot be fixed. Or one and the same bug might get logged 4 times only with a different platform specified. Solve like issues before you start!
- *Focus on "postponed" and "won't-fix" bugs in previous versions*, especially ones with low severity and priority: do they need to be logged?
- *In specific test areas consider testing only target files.* When you perform the same checks repeatedly on many builds and cannot automate this, you might focus on testing the target, rather than ill-conceived comparisons with the source. For example, you can apply this approach in the testing of a web learning course.

Prep tasks	preparation before project; OS setups and config; build setups; additional installations such as tools; server labs setup and config
Testing tasks	carrying out tests; running scripts
Admin. tasks	logging results; analyse automation logs; bug verification; bug closure; investigation on potential or ambiguous bugs; build/component problem solving; queries

Fig. 1: Three test tracking categories for advanced activity-based projects

Some of this might help you avoid logging irrelevant bugs. Be aware that you may have a project or a language where your customer sees relatively few bugs logged, and therefore becomes concerned and suspicious. They will ask you, for example, why there are so few bugs, whether your team is on track and if not why not. There really could be something wrong, so do dig down and be ready to explain.

Your test engineers are undoubtedly smart, but it always seems right to teach them to rely on their common sense when in doubt. In case of behaviour that you saw the first time around it is better to log the problem once and consult with the customer on how to deal with further occurrences.

Tracking the Nose on Your Face

What is the optimal level of tracking on a test project? Let us take the viewpoint of a vendor test manager who wants to keep the project budget under control, and who is ready to answer the following customer question anytime: "What were these hours spent on here and why should I approve it?"

There is a Czech saying that goes, "To get a lot of music for little money." Most customers today require excellent quality under very tight budgets. The majority of test projects today are activity-based, as opposed to resource-based, and majority of test management is thus with their business partner. The difference is that you get volumes in testing hours for an activity-based project, while for resource based your customer would say, "I need 20 people for three months to test my new release," and will pay for those.

There are several risks to consider when the test PM decides how to set up the project tracking:

- *Despite good intentions for transparency, tracking standards may get complicated or garbled and do more harm than good.*
- *Activity types are easy to follow but they don't give you the information you want* - tracking does not allow for good reports and budget control.

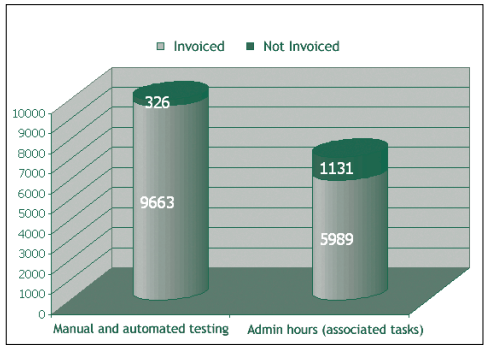


Fig. 2: High level test project results: Tracking systems should allow you to go deeper into the task level and unveil inefficiencies.

- *There is no tracking* and you never know where have all the hours gone.
- *You do not report what you spend* despite a tracking process in place.

So What Should You Track and How?

First of all, be sure your tracking will work with the structure of your budget and volumes from the customer. If the customer gives you volumes for preparation, manual testing, automated testing and final testpasses, consider tracking these task areas. Be aware that the tracking process will be each test engineer's daily routine. Fig. 1 shows three test tracking categories for advanced activity-based projects. It is very useful in the beginning, when you need to benchmark many tasks and verify estimates. Later in the project you can, to a large degree, discard this tracking and use higher level means. Be reasonable and keep this a process with a purpose that is understood by all, otherwise it is a waste of time.

It is worthwhile to introduce something like an invoicing checkbox in addition to the categories, specifying which tasks can be invoiced to the customer and which can not. Tracking activity groups allows you to provide precise reports and see where you are with estimates and actuals (See Fig. 2). Moreover, post-mortem analysis and evaluation of this data is priceless — you will see clearly what went wrong and where to make changes next time.

Orchestrating and Outsourcing Multiple Testing Teams

Your testing department has probably experienced something like this: You win a proposal, a great opportunity that you and the team have spent weeks on. The actual project takes 2 months to plan and prepare, but, suddenly, your close and trusting customer suddenly requires you to start in two weeks! It looks like you should join capacities to build a team of fifty test engineers for a 6 month project in no time and make sure it is one team that works. In your mind, the risk analysis looks beastly. Splitting the teams and locations could do more harm than good, especially in software testing and engineering. After all, you have been highlighting the value-add through team centralisation during the entire sales cycle. The list of inefficiencies you could suffer starts rolling out before your eyes. Chain handoffs, missing and lost information, triplicated preparation effort, duplicate, disparate quality standards, and low productivity, not to speak of problems with daily throughput, which is key for this project. These could be the potential results of testing on several sites simultaneously. But you do have capable staff at the central and branch offices, reliable testing partners, and a backup resource solution in place with 10 trained test engineers which you can have set-up and running in one week, don't you? Is it not a scary experiment to split the testing team in three on one project? What do you lose and what do you gain? What should you do to create a reliable and dynamic model with three parties involved? And how would the project run in this setup? The answer might be outsourced testing.

Let us assume there are 15 languages to test and a less reliable build plan which makes it hard to predict the weekly workload. Your solution is to create three main test locations; one at your Asian branch, one at your business partner in your city and one in your headquarters.

Prerequisites for a Successful Virtual Team:

- Technology, infrastructure and process must be seamless. If your branch offices and partners use the same hardware, have appropriate file transfer and connectivity, are familiar with your shared database tools and implement your test processes, then half of the problems are solved.
- An excerpt from a checklist unified for all locations could look like this:

Test machine: DELL GX260
Upload/download internet bandwidth: 4Mbit
File transfer: customer replication server X
Backup transfer: Secure FTP
Servers: Exchange 2000 SP1
Project sharepoint: http://extranet.companysite.com/project
Project knowledgebase

Obviously it is best to play an easy tune first and verify these processes on a small-scope project. Then you can be confident to play the symphony.

Efficient team model:

■ *Maintain a single point-of-contact with the customer*, and have experienced local test leads on each production site.

■ *Be sure there is only one project manager*; the production sites on large projects usually have a local PM.

■ *Communication and information sharing:* Be sure the testkit is available to all and is shared from one location together with the testplan. If not provided by the customer, create your own and share it on the project extranet portal. IP phone, Netmeeting, instant messenger and telephone are your daily means. Extranet online reporting, tracking and assignment tools are more than useful (See Fig. 3).

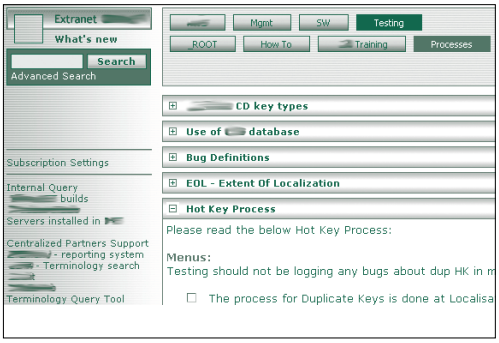


Fig. 3: Example of a project extranet

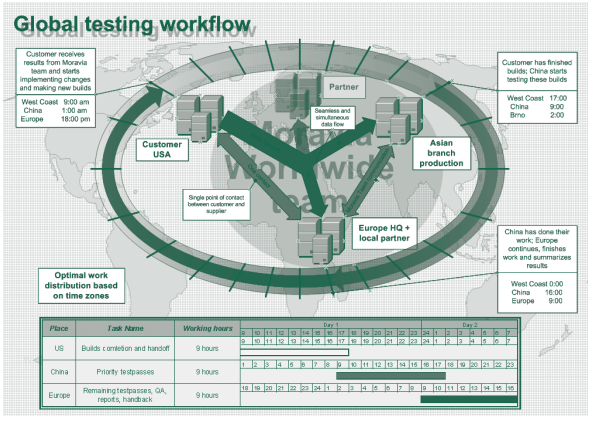


Fig. 4: Making the most of global workflow

Making Multiple Locations an Advantage:

The Earth is round and 3 x 8 = 24: While making sure your working hours overlap, you can do more in one day by setting up the appropriate workflow (See Fig. 4). The customer hands off to a single vendor contact and the Asian team start right after handoff. Then the European teams take over. At the end of the European day you can report the overall results to the customer. This solution is great when you need fast tested content turnaround. It relies on precise completion tracking and online reports within vendor production sites.

■ *Centralising, sharing, and work distribution experience:* The centralisation does not get sacrificed. It is there at the level of test leads that are responsible globally. The processes are set as if the team was sitting just in 3 rooms of one building. The right team needs to do the right stuff — in this case it would be the double-byte languages that get tested in Asia. One production site would usually specialise in demanding server testing. However, Asian teams should have the same capabilities of testing Latin character languages. This model allows for a more dynamic resource planning within the whole testing group.

■ *Return on Investment (ROI):* An added benefit of this model is that it allows for better strategic positioning due to performing testing in price-attractive regions.

We test engineers and (test) managers should pretend potential risks are a little more serious than they appear. Not waving them off can provide you with a probing analysis and will give you more control over your project results. In this short series we have tried to show a few real-life examples, tips and methods that I believe help better identify and clarify the stakes of project preparation, test planning, tracking and team models. Be a little paranoid at sensing issues; then reasonable, QA-conscious and aware of customer value while solving them. ■

Roman Civin is a senior manager with Moravia IT and oversees the testing division in Moravia's headquarters in the Czech Republic. He is responsible for European and Asian functional and linguistic testing services including localisation QA, automation and internationalisation. He can be reached at RomanC@moravia-it.com

Can Productivity Gains Continue?

Tex Texin

The release of Unicode 4.0 this year reminded me of how far we have come since the Unicode Consortium was formed in 1991. Thanks to the Unicode Standard and the Unicode Consortium which develops it, the software industry has undergone tremendous improvement in both developer productivity and typographic quality, for products supporting multiple languages. Will the next decade bring similar improvements?

Pre-Unicode

Back in 1991, there was no World Wide Web. Even e-mail was limited in its reach. It certainly wasn't the global phenomenon we know today. Information on character encodings was sparse. In the instances where it existed, it was often available only in the native language. Of the information that was in English or translated, it often had mistakes in critical areas. Ken Lunde's first book, Understanding Japanese Information Processing, which became the bible on Japanese encodings for many of us, was not published until 1993.

As there was no international encoding standard, we were left with a plethora of character repertoires and encodings being generated by national governments and hardware and software manufacturers.

Technology companies were just beginning to recognise that they needed to provide information about their different localised products to developers outside of their local markets. Most companies simply did not have the infrastructure to translate what were most likely internal documents and make them suitable for publication. To get information about a localised product, you would have to call overseas to the company's regional support office. Of course, the overseas office often could not accommodate English speakers, or the expense of international mailings.

To be fair, Unicode has some wrinkles of its own.

One of the biggest challenges was identifying "language experts". Often, resources provided by regional sales partners turned out not to be helpful. Details on language rules, sorting, etc. that they provided were incorrect. It is just not the kind of information that the layperson knows well enough to define with sufficient precision for computer algorithms. Simply communicating with these "experts" was problematic. They were often in a time zone on the other side of the world. Their command of English was usually not good, and the terse faxes could be ambiguous and filled with cultural mis-cues.

Many products implemented their own encoding and language support to ensure cross-platform consistency. For each combination of language and encoding, data tables identified characters as alphabetic, numeric, multi-byte, double-width and other types, and defined upper-lower case relationships, collation rules, etc. For example, one product's language-encoding tables had:

- 40 character properties
- 45 case rules
- 40 word breaks
- 100 collations
- 250 encoding conversions

For the simplest languages, research and implementation were quick. We could mark up an 8-bit code page, create tables and test cases in a week. However, not infrequently, languages would introduce new requirements to the software. Examples include moving beyond 8-bit encodings to double and quadruple-byte encodings, specialised keyboards

and input methods, bidirectional languages, collations like Thai that swap certain characters, etc. Often the new wrinkles would require not just table format and algorithm changes, but API changes and significant rewrites of the application which would be measured in man-years rather than weeks. The effort could be so large that it simply was not profitable to do so.

Today's Unicode

A decade later, the landscape is completely different. Thanks to the Unicode Standard, developers have clear guidelines for implementing text support for most languages. You can define an API once, and be confident it will meet the needs of nearly any language. Today, most research effort has been eliminated, as Unicode publishes character property and other data. Some customisation can be required, for example to tailor either collations or encoding conversions, but where these are not provided by the Consortium, they are often available from other resources, for example IBM's ICU library. Identifying language experts to answer your implementation questions is no longer hit-or-miss, as the Consortium runs mailing lists where numerous experts congregate and are willing to provide clear answers. As a result of the Unicode Consortium's comprehensive guidance, applications today can be more culturally correct and feature-rich than possible 10 years ago at far lower cost.

The productivity gains of the last decade came about because Unicode was more than a superset of characters.

To be fair, Unicode has some wrinkles of its own. Unicode 4.0 has more than 96,000 characters. The simple tables of 8-bit code pages are replaced with more complex access methods that can be efficient in storing and retrieving character property information. Having so many characters, there are also some redundancies, and so Unicode data may need to be "normalised" to perform proper comparisons. But these issues are minor programming problems considering the gain in using a single standardised approach for any and all languages and not having to rewrite API and applications every couple of years.

Unicode TNG (The Next Generation)

Which brings me to my question: Will Unicode provide similar gains in productivity over the next 10 years, and if so how? The question is important in times of tight budgets. Companies are reevaluating which standards groups they should sponsor and invest in. Individuals also consider which activities best preserve their future employment possibilities. Many naïvely believe that the work of the Consortium is sufficiently complete, since languages needed by the majority of business applications are supported today and Unicode functionality is provided by most operating systems.

The Consortium has not published a roadmap for future versions of the standard, other than outlining potential character and script additions. Members of course have access to draft reports and internal mail lists. Here is my personal divination of the Unicode tea leaves. (Note I am not representing the consortium, these are just my opinions.)

The productivity gains of the last decade came about because Unicode was more than a superset of characters. The collection of associated properties and algorithms, and especially the guidance provided on their usage was critical to reducing development and integration efforts. Here are some areas where further gains can be made.

Improving Multilingual Utilisation

Most Unicode applications today still maintain a monolingual mindset, expecting users to work predominantly with one language at a time. In truth, it is not always clear how to work with multilingual data. For example, when text is mixed in Latin, Cyrillic, Ideographic, and Thai scripts, what is the correct ordering of the scripts? There are also numerous and well-documented typographic problems with mixing text from different scripts.

Another problem is how to express search, replace, and filter across multilingual data. For example, searching "A-Z" often signifies "match all characters". In a multilingual environment, "A-Z" represents only a small percentage of the possible characters. A technical report providing guidelines for regular expressions may address this. As most programming languages make use of regular expressions, this is likely to have a very large impact on the industry. Multilingual expressions will also be a source of powerful new features and capabilities for end-users.

Conformance

When a vendor claims to support the Unicode Standard, what does that mean? Do they correctly and precisely implement the algorithms? Do they meet the standard's conformance criteria? Can vendors integrate Unicode-based applications and expect proper behavior across the Unicode domain?

The Consortium is identifying how applications can be assessed for conformance to the standard. Today, developers must design and create their own suite of tests and test data to verify the Unicode behavior of their software. This can be a substantial effort. The Consortium can greatly reduce the size of these efforts and improve the interoperability of Unicode applications by documenting test designs, establishing assessment criteria and making available appropriate test data.

Rich-Text, Verticals and Creativity

The Unicode Standard is designed for plain text. When employed in a higher level protocol or format, tradeoffs must be made between the use of Unicode's facilities and those available from the higher level protocol. For example, the Unicode Consortium and the World Wide Web Consortium have jointly defined a technical report that prescribes how to use Unicode in the context of XML and markup languages and the document is a good model for other protocols.

Mathematics uses characters from several scripts and small stylistic differences (bold, italic, superscript, etc.) are assigned significantly different semantics. It is therefore important to have industry agreement on the appropriate Unicode character and styling to use to represent mathematical symbols. Another technical report will provide guidance on using Unicode in mathematics. It is possible other disciplines would benefit from their own usage guidelines.

End-users and developers are also just beginning to explore the capabilities of combining characters and the possibilities for creating and using new symbols dynamically. Scientists and mathematicians do this frequently. It is sometimes evidenced in modern marketing materials. Font technologies and ubiquitous Unicode can make this expressiveness available to everyone.

Generalising these events, the Consortium might provide further guidance in the application of Unicode within other higher level formats, perhaps along vertical lines, and in using its combinatorial power.

More Scripts

Character repertoires continue to be defined by governments and vendors. Recent examples include the Hong Kong Supplementary Character Set (HKSCS) and the Chinese GB18030 encoding. The demand for continued character expansion is real. For example, the HKSCS includes characters used in street names used in Hong Kong. New technologies (e.g. cell phones, PDAs) and new styles of user interface often require new symbols, some of which may be incorporated into language and require encoding. Unicode must keep up with new character sets to remain a universal encoding.

Conclusions

The Unicode Consortium has created a single character standard that can be used universally and greatly reduces the engineering costs of international applications. The traditional work of incorporating scripts will continue. Applications using rich-text environments and those supporting particular disciplines can be made more efficient. Assessing applications for conformance to the standard is a sizable task where significant productivity improvements can be had. Users and application developers have much to learn about working with multilingual data and innovative multilingual operators have yet to be defined. The Consortium has another decade of interesting challenges in front of it! ■

Tex Texin founder, chief architect and Xen Master of XenCraft, a consulting and training company. XenCraft is a member of the Unicode Consortium, and Tex has been using Unicode to internationalise products since 1993. He can be reached at Tex@XenCraft.com

Localisation Focus Subscription

From September 2003 *Localisation Focus* will become FREE for members of The Institute of Localisation Professionals-TILP (www.tilponline.org). For non-TILP members there will be a cover charge of €7.50/$7.50 per issue or €20.00/$20.00 per year.

However, if you:
complete an LttN course (www.localisation.ie/LttNWeb/index.htm) in 2003, you can get issues 3 and 4 2003 — FREE!
supply the LRC with the details of 10 of your colleagues, you get issues 3 and 4 2003 AND an annual subscription for 2004 — FREE! Your colleagues will get issues 3 and 4 2003 — FREE!
subscribe now for 2004, you will get all remaining 2003 issues — FREE!

If you would like to subscribe to *Localisation Focus* OR apply for TILP Associate Membership, just fill out the subscription form and mail it to the address below, or fax it to the following number: Fax: +353-61-202734

Localisation Research Centre (LRC),
Department of Computer Science and Information Systems (CSIS),
University of Limerick,
Ireland.
Tel: + 353-61-202881

Online subscription forms are also available at www.localisation.ie
If you have any queries please email lrc@ul.ie

2004 Subscription Form

Please tick the appropriate box below and fill in your details.

For *Localisation Focus* subscriptions, please send a cheque with the appropriate amount made payable to: Localisation Research Centre.

For TILP Associate Membership, please send a cheque with the appropriate amount made payable to: The Institute of Localisation Professionals.

2004

4 issues, 2004, €20.00/$20.00
FREE *Localisation Focus* for TILP Members
1 TILP 2003 Associate Membership, up to December 2003, €25/$25
1 TILP 2004 Associate Membership, €50/$50

*Your Details

Name: _____
Freelancer (Y/N): _____
Company: _____
Position: _____
Email: _____

Address: _____
City: _____
Postal Code: _____
Country: _____
Tel No: _____

*The LRC will not give your details to other parties without your prior consent.

Supply the LRC with the details of 10 of your colleagues and get issues 3 and 4 of 2003 for **FREE**. In addition, you will also receive a **FREE** subscription to *Localisation Focus* for 2004. Your colleagues will receive issues 3 and 4 of 2003 — **FREE**!

Supply the LRC with the details of 10 of your colleagues and you can get issues 3 and 4 2003 AND an annual subscription for 2004-**FREE!** Those colleagues mentioned will get issues 3 and 4 2003-**FREE!**

1.Name: _____ Address: _____
Freelancer (Y/N): _____ City: _____
Company: _____ Postal Code: _____
Position: _____ Country: _____
Email:_____ Tel. No: _____

2.Name: _____ Address: _____
Freelancer (Y/N): _____ City: _____
Company: _____ Postal Code: _____
Position: _____ Country: _____
Email _____ Tel. No: _____

3.Name: _____ Address: _____
Freelancer (Y/N): _____ City: _____
Company: _____ Postal Code: _____
Position: _____ Country: _____
Email:_____ Tel. No: _____

4.Name: _____ Address: _____
Freelancer (Y/N): _____ City: _____
Company: _____ Postal Code: _____
Position: _____ Country: _____
Email:_____ Tel. No: _____

5.Name: _____ Address: _____
Freelancer (Y/N): _____ City: _____
Company: _____ Postal Code: _____
Position: _____ Country: _____
Email:_____ Tel. No: _____

6.Name: _____ Address: _____
Freelancer (Y/N): _____ City: _____
Company: _____ Postal Code: _____
Position: _____ Country: _____
Email:_____ Tel. No: _____

7.Name: _____ Address: _____
Freelancer (Y/N): _____ City: _____
Company: _____ Postal Code: _____
Position: _____ Country: _____
Email:_____ Tel. No: _____

8.Name: _____ Address: _____
Freelancer (Y/N): _____ City: _____
Company: _____ Postal Code: _____
Position: _____ Country: _____
Email:_____ Tel. No: _____

9.Name: _____ Address: _____
Freelancer (Y/N): _____ City: _____
Company: _____ Postal Code: _____
Position: _____ Country: _____
Email: _____ Tel. No: _____

10.Name:_____ Address: _____
Freelancer (Y/N): _____ City: _____
Company: _____ Postal Code: _____
Position: _____ Country: _____
Email:_____ Tel. No: _____