[*Translating* and *the Computer 20*. Proceedings from Aslib conference, 12 & 13 November 1998]

Translation in the Next Century: A Future Vision

Jens Thomas Lück, President & CEO
Logos Corporation

There are two words that I believe will characterize the translation industry as we look ahead at the 21st century. The first of these is the word "digital." Like the generation and handling of virtually all other information in the corporate environment, translation in the 21st century will be accomplished in a digital manner. Human skill of course will always be at the heart of this translation effort, but human skill will be *digitally* leveraged, to an extent I suggest that will be well beyond anything we have seen so far. We can expect to witness, I think, a symbiosis of human skill and digital processes that will effectively transform translation as we know it today.

This notion of symbiosis leads to the next word that will characterize translation as we look ahead to the coming century—that's the word "integration." Not only integration of human skill and digital processes, but integration, deep integration, within those digital processes themselves, so that a single, seamless digital carpet, so to speak, covers the entire translation process.

This digital integration is going to happen because it has to. We already see signs that traditional processes will not keep up with increasing demands for speed, quality control, and cost control such as are being placed upon the traditional methodology. We already see the translation industry changing in response to this pressure—accelerated growth, mergers, acquisitions, and above all, a scrambling for technology. For it is generally recognized that technology and technology alone is what will make it possible to address these pressures of time, cost and quality in effective ways.

Let me draw a brief picture for you of what this new, integrated, digital world of translation might look like in a concrete situation.

Company X decides one day it would be a good idea to introduce some order into its multilingual documentation process, beginning sensibly enough with the nomenclature the Company uses to describe its products. The Company hears about and acquires a brand new *multilingual documentation management system* called Rosetta 2000™. Among other things, Rosetta 2000™ has powerful terminology management facilities that serves Company X's immediate purpose regarding nomenclature quite nicely. This terminology management function combines parsing, statistical massaging and text alignment in a way that allows Company X to extract its terminology in multilingual form from all its past documentation still considered relevant. Rosetta 2000™ then stores all this nomenclature in an on-line multilingual terminology bank for use in each step of the documentation process in the future, both on the source and target side. Building this lexicon of course took some time and effort, since inconsistencies and non-standard usage had to be detected and weeded out. But now Company X has a firm handle on the language of its business, and that's going to pay off handsomely as I hope to be able to demonstrate in what follows.

Let's imagine for a moment now that Company X is developing a hot new product that it wants to release simultaneously in a dozen marketplaces, each entailing a different language. Competition is intense so time-to-market issues have critical bearing on market share. So that translation not be a bottleneck, Company X requires that translation be done in tandem with source document authoring. Company X can do this because Rosetta 2000™ has a very effective multilingual document management function that helps coordinate the work of authors and translators scattered all over the world. Thus, as soon as a source draft is completed, Rosetta 2000™ routes it to translators connected by corporate intranet all over the world. Thanks to Rosetta 2000™, the back and forth shuffling of text that becomes inevitable when authoring and translation are done in parallel all comes off without a hitch. Barring human glitches, the Company meets its tight deadlines with a document that is actually superior in quality to anything it did in the past and at substantially lower costs.

Where and how are these gains in time, cost and quality being effected?

Well, let's look at the process in more detail. When author Y of Company X gets an assignment to compose a document, or revise an existing document, he or she does so in the context of Rosetta 2000™'s authoring tool. This authoring tool does two things for the author. The first concerns *standard usage.* The author's text is dynamically processed against Company X's terminology bank to insure that only company approved nomenclature is used. If the author uses the term "mainframe," for example, and the Company-preferred standard is "server," Rosetta 2000™'s authoring tool will point this out to the author. Any new terminology the author introduces that is not already in the terminology bank will be captured, along with frequency statistics and sample contexts, and channeled to the responsible lexical standards specialist for review and signoff. These new source terms are then routed to translators over the intranet for their transfers, which then of course are stored in the terminology bank. Thus, when this author's text is ready to be translated, all the terms and all their transfers have been approved and placed on-line, available to everyone involved in the process.

The second function of Rosetta 2000™'s authoring tool concerns the translatability of the source document. Company X's management has come to realize that if they are translating a document into a dozen or more targets, then the correction of a single problem on the source side may well save twelve corrections on the target side. And lots of cost therefore. That consideration more than anything else seems to be driving the growing interest in authoring tools nowadays.

Translatability problems typically relate to text that is difficult or ambiguous to translate. Rosetta 2000™ has a tool designed to measure this called a "translatability indexer." Using a parser and certain statistical measures, it assesses each sentence of a text for complexity and any sentence found to exceed a certain complexity threshold gets flagged. Now it's well known that authors do not like authoring tools that try to tell them how to write. Rosetta 2000™ has a light touch in this regard. All Rosetta 2000™ does is flag the occasional sentence that it knows will cause problems on the translation side. Rosetta 2000™ can do that because it parses each sentence and knows where the problems are going to be. Heeding these sentence-specific warnings will clearly increase the useability of raw machine translation output, and lower the final cost of translation therefore. To be sure, not every sentence that parses poorly in MT is necessarily a bad sentence that needs to be fixed, but experimental testing has shown there's a high correlation, strong enough to make this sort of thing well worth while. Flagging overly complex sentences is worthwhile even where machine translation is not contemplated. Clear writing benefits the human translator no less than the machine.

Rosetta 2000™'s authoring tool also produces a translatability index for the document as a whole, on a scale of one to seven, based upon complexity factors such as sentence length, number of clauses, number of commas, parenthetical embedding and so on. Any translatability index reading below five is considered unacceptable by Company X's management, so the writers can be allowed to use their own judgment about writing style just so long as their writing style does not cause their texts to get an unacceptably low translatability index reading. Of course, the higher the translatability index value, the smoother the translation process at the target end, with commensurate benefits regarding cost. With such a measure in hand, it's now conceivable that management might want to reward authors whose writing style saves the company money.

We're at the point now where the author's drafts are ready to be translated. Before routing the drafts to the translators, however, Rosetta 2000™ creates its own provisional digital translation of the draft, drawn from its translation memory module and from its machine translation engine. The output of these processes of course is color-coded so that the translators, when they get this output up on their screens, can see by the color codes where the translation came from.

To be sure, not all texts are going to be revisions of something already stored in translation memory, and not all target languages are going to be covered by machine translation. This means that some texts must be translated entirely by hand. But even here, the digital resources of Rosetta 2000™ have something to offer. If a source text is being translated into a new language like Greek, let us say, for which Company X has no translation memory and no MT coverage, what will happen is this: assuming the terminology bank work has been done up front, as described earlier, the source text will be sent to the translator's screen with the Greek terminology placed in line right beneath the source language equivalent. The translator can then drag and drop these Greek target terms into the target text as it is being built. This sort of help can be very meaningful to the translator. Studies have shown that up to 70% of a translator's time can be spent on lexical research. Thus even in purely manual translation, Rosetta 2000™ offers substantial assistance.

The interesting thing about all this is that a single Rosetta 2000™ terminology bank, the one that Company X built at the beginning, now underlies all these translation processes, insuring uniformity no matter what the translation method, whether manual, translation memory, or machine translation. This is the benefit of the deep integration realized in Rosetta 2000™.

The text is now on translators' screens in different parts of the world. Two windows are open, one for the source language, one for the target version. When the translator scrolls in the one, the other scrolls with it. In the target window, color coding indicates the source of the translation, whether from TM, fuzzy TM, MT. The latter two, of course, the translator has to review and revise or possibly even discard and re-do manually.

The question is how effective will this be. The numbers are already well known and well established: translation memory users acknowledge that savings of 30% and higher are not untypical for documents that must undergo periodic revision; MT users for their part have fairly consistently acknowledged savings of 50%. Indeed, according to some users, the savings can rise to 70% if preparatory lexical work is done and the source document is well written. In the context of Rosetta 2000™, then, assuming things are being done correctly right from the outset, Company X can anticipate very substantial savings in its translation costs.

And to these cost savings we must add the savings in time and the gains in quality control, which have cost implications in their own right.

Rosetta 2000™ is not finished yet. As the post-editor corrects the translations, Rosetta 2000™ captures certain of these edits. Any edited sentence of course is captured as such and stored in sentence memory in both source and target form. But Rosetta 2000™ is more powerful than existing translation memories and also reaches *inside* the sentence, capturing for example any edits made to noun phrases. These edited phrases are stored and semi-automatically applied to further instances of that phrase in the present document or in any future document. In effect, the post-editor has to make such a correction only once for it to be applied where appropriate in the current or any subsequent document. Rosetta 2000™ is even sophisticated enough to be able to adjust case endings for noun phrases in languages like German where this may be necessary. There are limits to this, of course, but for many situations, what I have described will work quite nicely.

Rosetta 2000™ is able to support translation memory at the sub-sentence level because of the depth of integration effected between the TM and MT modules of Rosetta 2000™. In sum, the parse information obtained about each sentence of the text during machine translation is preserved and made available to the post-editing environment where translation memory capture takes place. It should be noted that sub-sentence elements that are edited at the local translator station can also be routed back to Company X's central terminology bank for review. Once approved, they would thereafter be applied to all subsequent translations fully automatically.

If you think about it, this automatic feedback from the post-editor's work to the knowledge base of Rosetta 2000™ is quite revolutionary, for it means that the terminology bank itself could conceivably be built and maintained as a direct by-product of system usage. It's revolutionary because the Rosetta 2000™ knowledge base literally grows as it is used, becomes smarter the more translators work with the system. This is machine learning and we can expect to see much more of it in the future. In fact, a crude experimental prototype of what I have just described already exists at Logos' technology center. The experiment more than established feasibility for this sort of thing.

I should say something about the system environment for Rosetta 2000™. It's a network-enabled client-server system. The server will run on a UNIX workstation under Sun operating systems or on any PC under NT and any successors of these operating systems. The client, being implemented in Java, will run on any desktop PC regardless of the operating system. All the knowledge bases, the terminology bank, translation memory, and MT itself, will be network-accessible and thus always up to date and shareable by all. The knowledge store will be a relational model, so that management can access, distribute, re-use and re-configure its knowledge base asset as needs dictate.

Some of you may be thinking that virtually all of Rosetta 2000™'s features are available now in one form or another. And so it is, or virtually so at least, which is the good news. That means that this vision I have been presenting is not sheer fantasy but real and about to happen. What's still missing however is the deep, seamless integration that makes all of this functionality readily available to the user. Lack of such integration has been the biggest problem with MT in the past, for example. Offering something like machine translation in the past, without concern for how it integrates into the rest of the documentation processes, was like offering someone a beached whale. What do you do with it? The same with the lack of deep integration between TM and MT, where the user was obliged to maintain two separate lexicons. The same may be said for authoring tools that seek to constrain a writer's style without being able to justify why one style was really better than another. In Rosetta 2000™, sentences are flagged because the authoring system has found them to cause parsing errors. You have to have a strong parser to do that effectively, to be sure, but Rosetta 2000™ will have a strong parser. As for the writer, this is a much more focused sort of constraint, one that few authors will want to argue with.

The nice thing about a system like Rosetta 2000™ is that there's something in it for everyone, manager, author and translator alike, no matter what the responsibility or prejudice. Indeed, the walls separating authors from translators, manual translation from digital processes, and translation memory from machine translation all tend to dissolve. Translator workbenches have been aiming at just such a world for some time now. The Rosetta 2000™ vision thus can be seen as the ideal workbench, not only to translation but extended now to a Company's entire documentation process from beginning to end.

Something like Rosetta 2000™ will happen because it has to. It will happen because the technology is virtually all there that will allow it to happen. What's needed now is for someone to put it together in a seamless fashion and make it available to the world. It will take considerable money and effort to do this. There's a great deal of non-trivial technology involved and it is unlikely that any one company by itself can muster all the technology needed. But we can be certain that it will happen and probably happen perhaps sooner than we think. Indeed it is already happening. As they say, necessity is the mother of invention, and it's becoming abundantly clear that the times are calling for something just along the lines of Rosetta 2000™.