[*Translating and the Computer 24*: proceedings of the…International Conference…21-22 November 2002, London (Aslib, 2002)]

THE FULL-TEXT MULTILINGUAL CORPUS: BREAKING THE TRANSLATION MEMORY BOTTLENECK

Daniel Gervais
MultiCorpora R&D, Inc.
Canada

dgervais@multicorpora.com

INTRODUCTION

Driven by fast-paced global competition where the time-to-market of new products, services and communications into multiple languages and cultures is mission-critical, organizations are increasingly demanding translation services that provide faster turnaround while maintaining the highest level of quality. A key driver behind the need for speed and quality is the ongoing explosion of web-based content and the related expectations of content freshness and quality. Operating in a competitive and typically fixed-price environment, translation service providers need to respond with significant gains in translator productivity while continuously improving translation quality. Also, translators and terminologists do not work in isolation - they are members of a complex language management value chain that consists of monolingual and multilingual authors, reviewers, translators, terminologists, and content consumers who may reside in multiple organizations. In order to fully optimize language management, all of these participants must be able to seamlessly share language resources and collaborate in real-time.

For many years now, Computer-Aided Translation (CAT) tools have held the promise of productivity and quality gains for translators. Despite considerable hype by the early software vendors, the first generation of CAT tools, referred to as translation memory, has failed to deliver. The up-front investment to populate multilingual terminology and translation memory databanks has proven prohibitive for many service providers and the actual productivity gains realized have been insignificant except for a few, very specific types of content.

Simply stated, translators are skilled writers who must take information in one language and communicate it in a different language while maintaining the precise meaning, style and tone of the original communication. Achieving these objectives requires much creativity and linguistic skill - skills that computers lack. While machines might be capable of generating approximate language translations for the purposes of indicating the gist of a written passage, they will likely never replace human translators for communications where accuracy and quality are important. With this firmly in mind, a different computer-based approach to helping translators has been developed and the industry is finally beginning to realize significant gains in translation productivity and quality.

This paper explores the promise of translation memory and the inherent limitations of traditional CAT tools. It then uses this as the foundation for introducing a different approach: the full-text multilingual corpus.

THE PROMISES OF TRANSLATION MEMORY

Conventional wisdom holds that there are few (some say no) original ideas or thoughts, just reflections or reinventions of the previous ideas of others. In translation work, this holds very true. Research clearly shows that, while complete sentences rarely recur from one document to the next, smaller expressions (mainly five words or less) recur very frequently. For many types of documents, recurring expressions account for over 50% of the words in the document. For technical documents, the level of coverage can exceed 75%. Even in the least-repetitive types of content, such as administrative communications and parliamentary debates, research shows that the level of expression repetition exceeds 25% of the words in a given text. In all cases, the majority of the repetition comes from expressions of 5 words or less. Complete sentence repetition, however, is almost non-existent in all types of content except documentation with a specific lifecycle, such as technical manuals for products that belong to a large family of similar products.

This high degree of repetition of expressions means that a large fraction of any new project has been previously translated many times. These previous expression translations exist in many accessible locations: in previous projects translated by the same translator, the work of other translators in the same organization, the work of other translators working for the same client, and publicly available translation work. The Internet alone provides convenient access to multilingual web sites that contain billions of words of translated text covering every topic. For example, a translator working on a project in the health field can reference hundreds of thousands of high quality translated expressions from the World Health Organization (WHO) tri-lingual web site.

Of the many previous translations for a given expression, some will not apply in a specific context because the terminology usage, style, and/or tone are inappropriate. However, many high-quality examples that fit the current context will also exist. If these previous translations of expressions could be effectively recycled and used as context-sensitive translation references, enormous gains in translator productivity and translation quality could be realized for all types of content.

TRADITIONAL CAT TOOLS HAVE FAILED TO DELIVER

Almost 10 years ago, Translation Memory (TM) tools began to appear on the market. The original concept was simple and has not changed significantly since then: human translator productivity can be increased and the consistency of translations can be improved if previously translated sentences can be stored in a database for later retrieval. This core assumption has proven to be flawed in several ways that we will explore later.

Despite considerable hype by the software vendors involved, TM-based CAT tools have found limited application. The only niche where they have gained a foothold is in the translation of technical product documentation (technical specifications, operating manuals, maintenance and support documentation, etc.). Moderate success has been attained when products share many characteristics (for example, multiple models of photocopiers within a product family) and when documentation requires frequent updating. This type of documentation is characterized by relatively frequent recurrences of whole sentences.

Even in this niche, however, there is a considerable gap between the productivity gains hyped by TM software vendors and the actual gains realized by translators. Vendors typically claim 75% to 100% or greater gains in productivity while independent reviewers cite gains ranging from 15% to 30% under ideal circumstances. The reason for the discrepancy is that CAT tool vendors use an unrealistic *theoretical* model to compute productivity gains rather than *actual* translator experience. Vendors typically use the example of updated product documentation within which 65% of the content is repeated whole sentences. Theoretically, the significant efficiencies would be realized for this repeated content, however, the reality is that much of the repetition in that type of document is in large sections of text, possibly entire chapters. Translators would not work through the entire document and re-translate large sections that have not changed, manually or otherwise. The project manager or translator would first identify the changed sections based on guidance from the author or using the standard "Document Compare" feature available in most word processors. They would then focus their translation efforts only on those new sections. Productivity gains are, as a result, negligible because the remaining content to be translated contains minimal whole sentence repetition.

Also, any actual productivity gains do not factor in the considerable effort to build the initial TM database before the actual translation work can begin - which for many projects completely offsets any productivity gain. To further negate any gains that may be achieved, clients typically demand lower translation price rates when a TM is used, thereby voiding any economic advantages of the tool for the translator.

Outside of the niche of repetitive technical documents with specific lifecycles, the productivity gains from TM-based tools have proven to be insignificant since repetition of whole sentences is extremely rare in most types of content. In fact, research indicates that existing CAT technology is appropriate for less than 5% of all documents that require translation.

To better understand the lack of success of these tools, let us take a closer look at their inherent limitations.

The Problems with Translation Memory

The three biggest limitations of TM are:

- Dependence on whole sentence repetition
- Loss of context
- Building a TM database is prohibitively labor-intensive

These three technical shortcomings significantly limit the positive impact TM systems have on productivity and quality - in some cases actually degrading both productivity and quality. To make matters worse, many translation clients (companies or large agencies who subcontract) insist on lower translation price rates when a TM is provided - offsetting the economic benefit of any productivity gains.

Dependence on Whole Sentence Repetition

TM systems are essentially limited to exploiting whole sentence repetition in previous translations. Most TM tools also perform "fuzzy matching" where whole sentences are compared and possible matches are identified if some percentage of the sentence

matches exactly. Translators report that fuzzy matches under the 85% exact match level are useless since it takes longer to correct and complete the translation of the partially matched sentence than to translate the entire sentence manually. Research into the characteristics of different types of documents in different domains (industries, topic areas, etc.) clearly shows that whole sentences (and 85% or greater fuzzy matches) rarely repeat. The only exception to this are certain highly structured technical documents, mainly related to complex product families.

Overall, significant sentence repetition occurs in less than 5% of all documents that are translated. TM systems are therefore largely useless for administrative memos, speech or meeting transcripts, marketing content, and most types of corporate studies and reports. The globalization of corporate web sites is another area where the vast majority of the content is not repetitive at the sentence level. While whole sentences almost never repeat, recurring expressions less than 5 words long account for the majority of the text volume in all types of content. Since TM systems do not effectively address short recurring expressions, they are unable to significantly boost productivity.

Loss of Context

Solving difficult translation problems is time-consuming due to the need for manual look-up of style and usage references, particularly for junior translators or translators new to a domain or particular client. Usage and style context is critical to making sound translation decisions.

Since TM systems maintain a database of isolated sentences, they lose the surrounding context within which the original sentence was used. The lack of style and usage context results in additional time-consuming translation review and editorial rework because translations built from isolated sentences are more likely to contain inconsistencies or errors. Further, by automatically "pre-translating", TM systems blindly reuse sentences that might not fit the context of the new project, resulting in poor-quality translations. To improve quality, the translator and others must spend extra time and effort to review, edit and correct - resulting in lost productivity.

Building a TM Database is Prohibitively Labor-intensive

Building a TM database is a tedious labor-intensive process of ensuring that 100% of the sentences in legacy documents are perfectly aligned with their corresponding translations. With the best alignment tools available, the process of creating a TM takes about 3,000 words per hour of translator time. Some alignment tools take up to twice that time. At those rates, building a sufficiently large TM database takes weeks or months of effort - making it prohibitively time-consuming in the competitive, low-margin business of translation services.

The millions of words of potentially valuable previous translations (internal and external sources) that exist around any given client or topic simply can never, in any practical sense, be fully exploited by the TM approach.

Some TM software vendors now use the term "corpus" to describe their TM database. By any name, a TM is a database of isolated sentences and their previously translated equivalents - not a full-text indexed and aligned body of multilingual content that provides context and the opportunity to find and reuse phrases and expressions below

the sentence level. They also suggest that users do not necessarily need to invest the effort to obtain perfect TM alignment prior to use, however, translators realize that with these tools, the effort required to review and correct translation errors resulting from misalignments would be more time-consuming than correcting these databases in the first place.

With TM-based systems, building large formal multilingual terminology banks (or responding to a translator's specific request for terminology clarification) is also an extremely time-consuming process, because terminologists lack the context-sensitive reference tools to rapidly research full-text legacy documents and create terminology records.

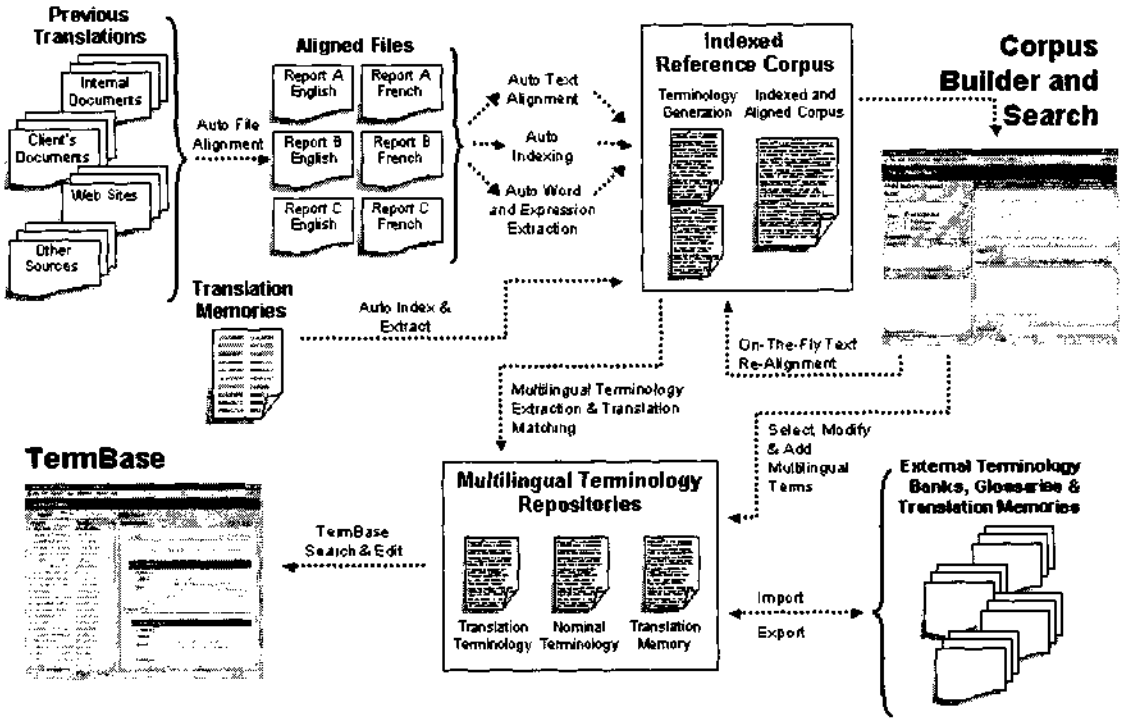
TIME FOR A NEW PERSPECTIVE: THE FULL-TEXT MULTILINGUAL CORPUS

Translation Memory tools were born and developed to address a very real problem: how to leverage previous translation efforts to realize significant gains in translator productivity and translation quality. MultiTrans™, based on full-text multilingual corpus technology, approaches the problem from a different perspective. It provides a solution that:

- Enables the rapid creation of vast reference pools of previous translation efforts
- Provides complete usage and style context for previous translations
- Effectively recycles translated expressions of any length, not just whole sentences.

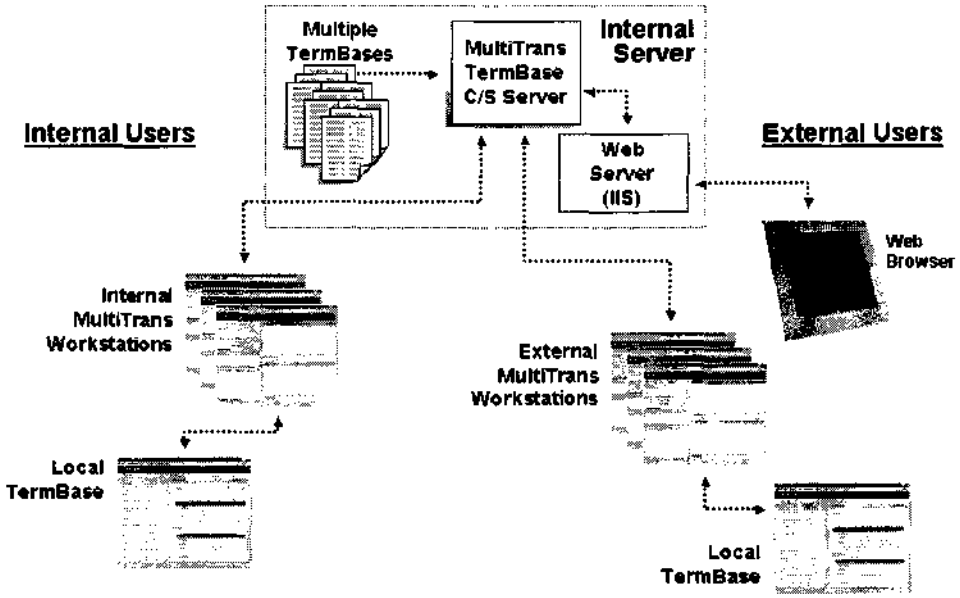
The MultiTrans Language Management Infrastructure

The integration of a full-text multilingual corpus and a comprehensive terminology management infrastructure sets MultiTrans apart from all other approaches to language management.



The diagram above identifies the two modules within MultiTrans that are used to build and manage reference material: the Corpus Builder and Search, and the TermBase. The Corpus Builder and Search capabilities enable the user to rapidly create and easily access large bodies of existing texts and their corresponding translated versions. These capabilities automatically align and index the full text and extract lists of recurring words and phrases. In addition, powerful search capabilities allow the translator to quickly identify all instances of a desired expression, and view them in their original and translated contexts. The TermBase provides comprehensive tools to extract and generate terminology from corpora or import terminology information from many other sources. Easy-to-use search and edit capabilities allow

the user to easily find and manage nominal terminology, translation terminology and terminology already contained in translation memories. Optional client-server technology allows for the deployment of all of this information so that language management tasks can be shared and collaborated across many users.



Furthermore, the infrastructure of MultiTrans allows it to be fully functional as a stand-alone product, or to be integrated with other language management systems already in place.

Instant Access to Millions of Words of High-Quality Translations

Rather than tediously building a database that contains isolated whole sentences and their previous translations, the corpus-based approach takes a vast collection of legacy documents and their previously translated sister documents and, using advanced search engine techniques, rapidly indexes all of the text. It also uses algorithms, similar to those used by TM tools, to align the translated text with the source text. A very large searchable corpus of previous translations can be built very rapidly - at a rate of approximately 50,000 words per minute on a low-end computer. A corpus of millions of words can be built in less than an hour and be ready for immediate use by translators. A TM database of comparable size would take years to build.

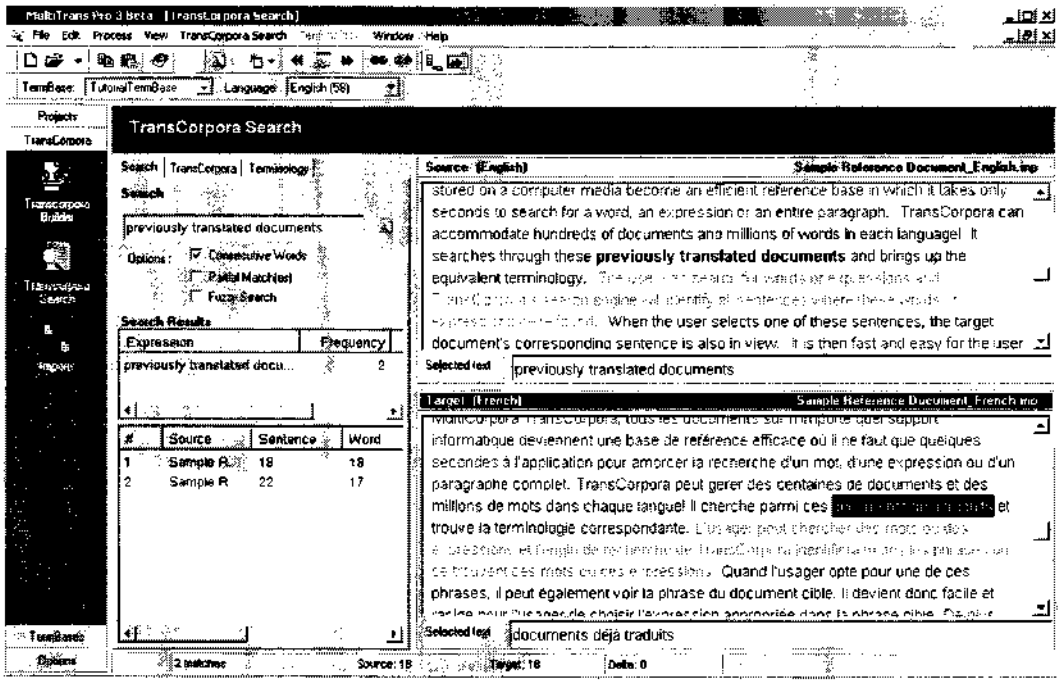
With this approach, it is possible for a translator to build a corpus in a few minutes by importing any relevant text in any relevant language. In addition to quickly building corpora from legacy "in-house" translation projects, any source of translated text can be easily exploited, including published web content. The potential benefit to translators of being able to easily reference web content is enormous. For example, for a translation project in the field of health care, a translator could quickly import a large quantity of relevant trilingual (English, French, Spanish) content from the World Health Organization web site and begin using it immediately in the translation process. Thousands of other web sites also provide examples of high-quality translations. A noteworthy example is Eur-Lex, an 11-language web site that contains vast amounts of content covering European legislation, treaties, case-law,

parliamentary debates, and documents of public interest. A translator working on texts related to food inspection, for example, could simply look up relevant European directives on Eur-Lex, click on the documents in the languages of interest and instantly create a searchable reference corpus.

The investment in existing Translation Memory databases is also fully leveraged since they can be automatically indexed and included in multilingual reference corpora where they can be searched and matched (exact and fuzzy) with new translation projects.

Translated Expressions of Any Length, In Their Full Context

A corpus user can perform a search of the entire corpus for an expression of any length, in any of the languages contained in the corpus. In less than one second, all of the instances of that expression in the entire corpus (in fact, multiple corpora can be searched simultaneously) are automatically found and retrieved, along with the aligned translated texts. The user can then select and view an instance of the expression and its aligned translation in a split screen view.



One half of the view displays all of the text of the document that contains the searched expression, automatically scrolled to the location of the found expression, which is color-highlighted for easy viewing. The other half of the view displays the complete corresponding translation text, scrolled to the aligned text segment, which is also highlighted. At a glance, the translator sees the expression and aligned translation in the contexts of their complete original documents. By providing context, the corpus acts as an extensive "by-example" dictionary for usage and style reference of terms and expressions.

On-The-Fly Alignment: No Up-Front Investment

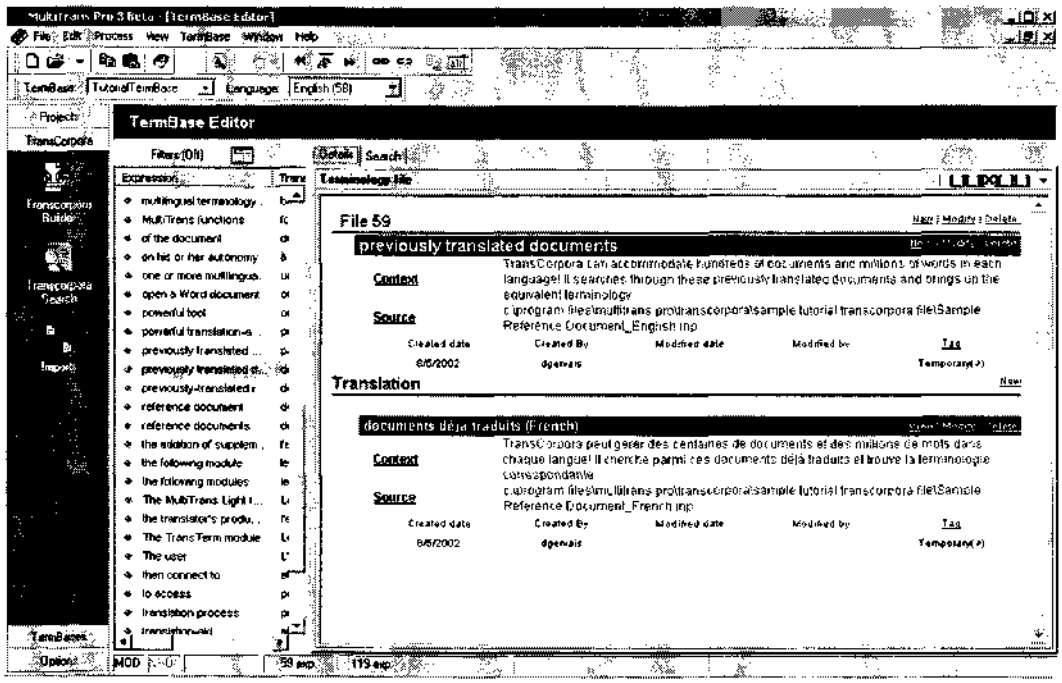
Despite using the most sophisticated alignment algorithms available, the two texts will not always be perfectly aligned. Misalignment may occur, for example when one sentence from the original language text has been split into multiple sentences in the

other language during the previous translation process. Since the text alignments are usually off by only a sentence or two, the user can see the problem at a glance in the full-text, split-screen views. A point-and-click environment allows the translator to easily correct the alignment on the fly, thereby continuously improving the corpus with use. Unlike TM databases, the corpus does not depend on perfect alignment, thus eliminating the need for time-consuming up-front verification before the system can be used.

Nominal Terminology, Translation Terminology and Sentence-Level Memory: The Best of All Worlds

If a found expression is a special term that requires formal terminological management, it can be simultaneously added to a multilingual terminology management repository. Terminology management is the domain of skilled terminologists who convert terms into nominal form and manage all of the surrounding information about the term. As with traditional TM systems, an integrated terminology management repository can be used to automatically pre-translate terms in a translation project.

Most recurring expressions (typically 5 words or less), however, are not part of the formal terminology of a subject but are simply sub-sentence units of reusable translation text. These expressions are sometimes referred to as "Translation Terminology" to distinguish them from terminology that is formally managed by terminologists. It is also important to note that this same repository structure allows the seamless incorporation of sentence-level translation memory databases, previously created with traditional TM tools, into the corpus-based translation process.

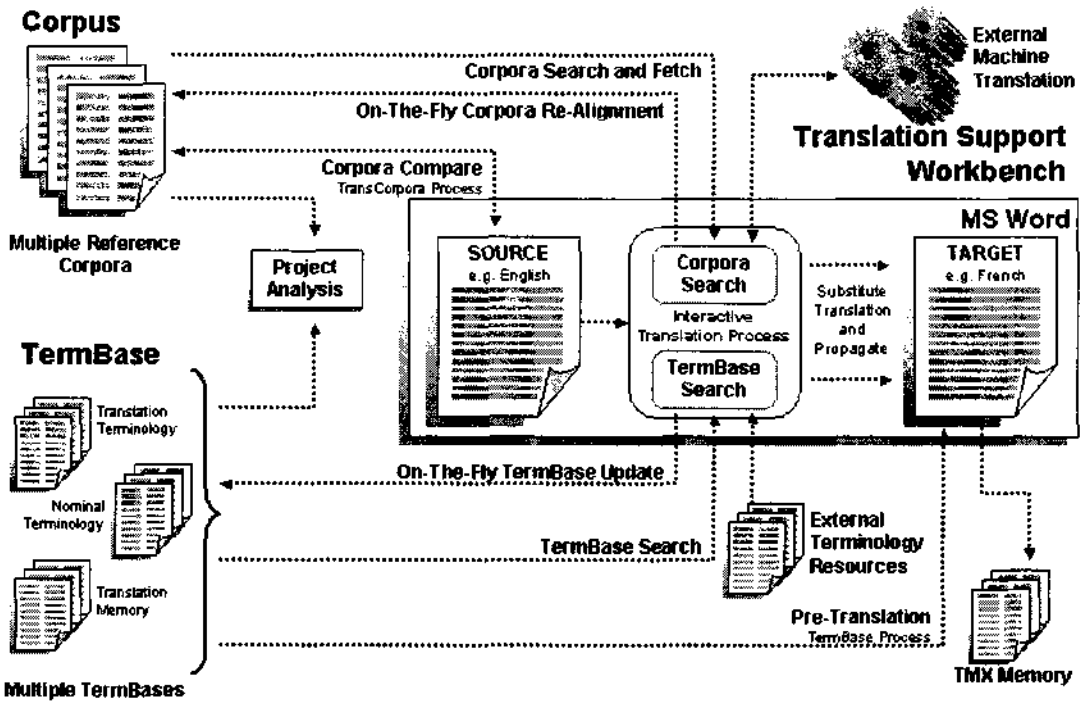


Since a translator has validated the translations in the terminology management system and the translation terminology repository, they can both be used for subsequent automatic pre-translation as well as manual search and retrieval.

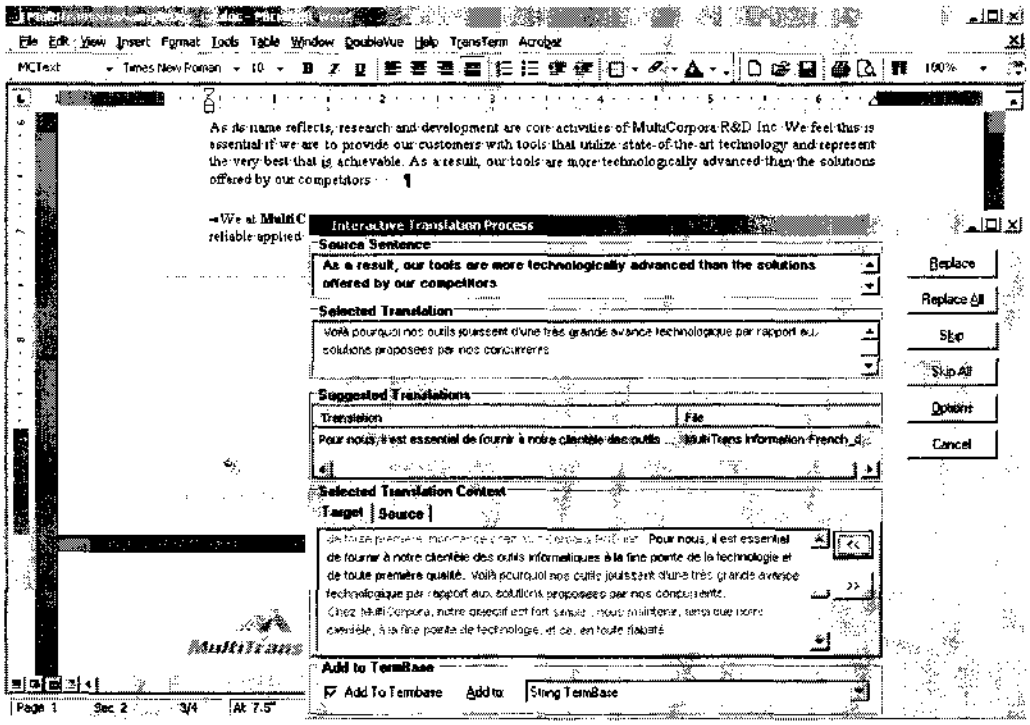
In addition to building formal and translation terminology repositories on the fly, the expression extraction and alignment tools of the corpus can be used to rapidly build these repositories independent of a translation project.

An Integrated Translation Support Environment

The MultiTrans Translation Support Workbench is an integrated environment within which translators seamlessly interact with multiple Corpora and TermBases to quickly complete new translation projects.



An interactive translation process enables translators to easily retrieve, review and insert corresponding translated terms and expressions from MultiTrans-generated Corpora and TermBases into new target translation documents. The Corpora also provides valuable usage and style context for expressions to help solve difficult translation problems. The Workbench can be further integrated into other translation resources, such as external terminology databases and machine translation systems.



Productivity and Quality That Improve With Age

By having easy and rapid access to multiple examples of the usage of an expression (of any length) and its previous translations in the context of the full texts to which they belong, translators can quickly solve difficult translation problems with high-quality solutions. Upfront TM database preparation work is eliminated, allowing translators to immediately begin exploiting vast reservoirs of legacy translations. And the interactive nature of the integrated corpus-based translation support environment allows the corpus, terminology management and translation terminology repositories to continuously improve with usage.

Benefits Across The Information Management Value Chain

We have focused on the translation process itself; however, the corpus-based approach offers benefits to many other participants in the information value chain.

The vast repository of expression usage and writing style also speeds and improves the quality of the monolingual new content authoring process by providing authors, reviewers and editors quick access to terminology banks and stylistic, definition and usage guidance from previous content.

Beyond the content-creation process, information consumers can also leverage these valuable resources to solve content-comprehension roadblocks by having easy access to dictionaries and usage examples. For example, if a knowledge worker referring to a

company document does not understand a key term, she can quickly look it up in the corpora and/or terminology repositories to see definitions and context-rich examples of meaning and usage. Corpora and terminology repositories become richer as more people interact with them, and synergies between translation activities and other information management processes can occur.

SUMMARY

Hundreds of millions of words of high-quality previous translations exist all around us, including vast resources on public multilingual web sites. Translation productivity and quality could be greatly enhanced if those previous translations could be effectively exploited.

Such has been the promise of Translation Memory (TM) systems; however, traditional TM-based approaches have simply failed to deliver significant benefits. At best, modest gains have been demonstrated in the niche of technical product documentation with specific lifecycles - which represents only about 5% of all documents that require translation. For the remaining 95% of content that must be translated, TM-based systems offer little or no advantage. Their failure stems from their dependence on whole sentence repetition, their loss of translation context and the prohibitive labor-intensity of building the initial TM databases.

Fortunately, a different approach to the translation-productivity problem, based on the concept of a searchable full-text multilingual corpus, delivers on the promise. The corpus-based approach enables the rapid creation of vast pools of previous translations, provides complete usage and style context for all translations and effectively recycles translations of expressions of any length, not just whole sentences.

The corpus-based approach provides clearly demonstrable superior productivity for all types of content, including descriptive texts that exhibit no whole sentence repetition. It also helps improve the quality of translations by providing comprehensive "by-example" usage and style references for all participants in the multilingual information-management value chain.