MultiTrans™ System Presentation
Translation Support and Language Management Solutions

Dan Gervais

MultiCorpora R&D Inc.
102-490 St-Joseph Blvd., Gatineau, Québec, Canada J8Y 3Y7
Telephone: +1 819.778.7070 www.multicorpora.com

Abstract

MultiTrans™ is a translation support and language management solution that is based on a multilingual full-text repository of previously translated content. It has helped global organizations and language-industry professionals to improve translation productivity and quality for all types of content. Unlike traditional translation memory tools, which are based on a database of isolated whole sentences, MultiTrans makes vast collections of legacy full-text translations searchable for text strings of any length in their full usage context. MultiTrans' interactive research agent automates and aggregates the search process, providing users with the most relevant information, maximizing language resource reuse.

1 Introduction

Vast bodies of high quality previously translated content may exist in many places: a translator's personal collection, a service provider's organization, clients' organizations, in many multilingual public web sites, and elsewhere. If effectively exploited, these untapped reservoirs of valuable translation knowledge would enable leaps in translation and terminology management productivity and quality for global organizations and language industry professionals.

Despite the obvious potential, Computer-Aided Translation (CAT) tools have failed to fully leverage these valuable multilingual assets. Translation Memory (TM) databases consisting of laboriously aligned, out-of-context full sentences have proven to be of only limited benefit and apply to only very specific applications, such as the translation of technical product update documentation. The three biggest limitations of TM, that in some cases actually degrade both productivity and quality, are:

1. a dependence on whole sentence repetition;
2. a loss of context; and
3. the building a TM database is labor-intensive.

Taking a fundamentally different approach to the problem, MultiTrans is a corpus-based translation support and language management solution that has been effectively used for all types of content in multiple domains, including the pharmaceutical industry, across all levels of government, multilateral organizations, and other business sectors worldwide.

MultiTrans:

- enables the rapid creation of a vast reference pool of previous translation efforts;
- provides complete usage and style context for previous translations; and
- effectively recycles translated expressions of any length, not just whole sentences.

2 Product Overview

MultiTrans is an integrated translation support and language management solution that provides productivity and quality gains to all of the participants in the multilingual information value chain, including writers, translators, terminologists, editors, reviewers and content consumers.

The complete MultiTrans solution consists of:

1. An indexed full-text multilingual reference corpus (a full-text body of previously translated content) and tools to build and search the corpus.
2. Terminology management tools, including a multilingual terminology repository and terminology extraction tools.
3. An integrated translation workbench that integrates with popular word processing environments to provide a variety of manual and automated capabilities to search multiple corpora and terminology repositories for examples of previous translations that are candidates for reuse.

4. A collaborative multi-user infrastructure for sharing language resources in real-time, including web browser access for the entire enterprise.

Since MultiTrans is UNICODE compliant and does not rely on linguistic knowledge, it can manage content in any written language, including bi-directional languages like Arabic and ideographic languages like Chinese.

3 A Full-Text Multilingual Repository of Previous Translations

MultiTrans enables users to easily build and maintain indexed repositories (corpora) of cross-referenced multilingual (or monolingual) content from legacy documents. The corpus-builder automatically aligns legacy translated document pairs based on user-defined file nomenclature. It then automatically extracts, indexes and aligns the complete text. The full-text indexing enables text strings of any length (words, sub-sentence expression, whole sentences or paragraphs) to be subsequently searched and retrieved. Statistical alignment algorithms establish links between the equivalent sentences in multiple language versions of the same content. When an expression of interest is found in one language within the corpus, the alignment link allows the corresponding previous translation to be also identified and retrieved.

When an expression and its aligned other-language version are retrieved, they are presented to the user in side-by-side scrollable windows showing each expression highlighted within its surrounding full text, providing the user with valuable usage and style context. Since automatic alignment algorithms cannot yield perfect results all of the time (for example, when a sentence in one language is split into two during translation), the full context views of MultiTrans allow a user to spot a misalignment at a glance and correct it on the fly. Because of this context and ability to correct the occasional misalignment as they work, translators can begin using a corpus within minutes of its automatic creation and the quality of the corpus actually improves over time with usage.

With MultiTrans, 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. Besides previous in-house translation projects, any additional sources of translated text can also be easily added to a corpus, 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.

4 Terminology Management

Large organizations face the challenge of maintaining a consistent corporate "language" in all of their communications, be they mono or multilingual. When multiple functions or geographic regions within an organization work together on efforts such as product development or marketing, terminology inconsistencies grow rapidly and can become a source of frequent rework and costly delays. The consistent use of proper terms (and the avoidance of the usage of undesirable terms) is also critical for communications clarity and accuracy and the reinforcement of an organization's global brand.

Since a MultiTrans corpus provides multiple examples of previous full-text translations, it serves as a valuable reference that reinforces a standard corporate language. MultiTrans also provides "terminology management" capabilities to complement the corpus with information on which special terms and translations have been specifically reviewed and approved for usage. In MultiTrans, managed terms can include nominal form terminology (as used by terminologists), any pre-approved translation of a word, expression, sentence or paragraph (sometimes called translation terminology), or translated sentences from Translation Memory files.

MultiTrans supports the terminology management process with a scalable database platform for storing and tracking comprehensive terminology management information, as well as a suite of integrated capabilities to:

- Enable users to easily capture new terminology during the authoring or translation process;

- Automatically extract terminology candidates and their corresponding translations from legacy documents;
- Provide tools and resources to support terminology research activities;
- Manage the approval status and lifecycle of terminology; and
- Make approved term translations available to translators for automated or manual inclusion in new translation projects via the integrated MultiTrans workbench.

Terminology extraction capabilities are included in the MultiTrans corpus-builder module. During the indexing and alignment process, the corpus-builder also extracts all of the recurring words and multi-word expressions and provides statistics on the frequency of occurrence of those elements. Algorithms analyze the terminology extraction and identify and retrieve probable corresponding translations from the corpus. The system then displays a list of possible translations per expression, and a terminologist or a translator can validate the suggestions. The system can extract over 60,000 expressions with their corresponding translations from a 6 million-word corpus in a matter of hours.

MultiTrans provides a simple point and click environment that allows users to rapidly review, approve and capture new terms. Direct navigation to all of the occurrences of the term in the full-text of the original documents that are contained in the corpus is also provided. By providing context, the corpus acts as an extensive “by-example” dictionary, usage and style reference for terms and expressions.

5 An Integrated Translation Workbench

While MultiTrans provides dedicated Windows and web-based interfaces to directly manage and search the multilingual corpus or terminology repository, it also provides a tightly integrated environment within Microsoft WORD or other popular editing environments (PowerPoint, WordPerfect, an HTML editor and others) where all of the language resources and search functions are available to a translator from within the documents that they are working on. Starting from the document to be translated (in its source language, say English), the translator interacts with MultiTrans functions to rapidly find examples of previous translations (in the target language, say French) of matching expressions, sentences and terms from the relevant multilingual reference corpora and terminology repositories. Previously translated segments that are selected for inclusion in the project are inserted into the document with a simple mouse click. MultiTrans provides a number of types of search capabilities to mine corpora and terminology repositories for examples of translations. An automatic pre-translation mode compares an entire new project in one batch operation to approved terminology repositories and automatically retrieves and inserts the corresponding translations. The translator can also manually select any expression in a new project and execute a search of the corpora and terminology repositories. More commonly, translators use a single automated and optimized multilingual search and comparison process that combines all search methods into one user action.

6 Automated and Optimized Multilingual Search and Comparison

Deploying a search and comparison algorithm, MultiTrans automatically compares an entire new source document to all of the open corpora and terminology repositories and proposes the set of matches that maximizes the suggested reuse of previous translation work. It does this by considering full sentence matches (exact and fuzzy), terminology matches, and thousands of expression matches covering all possible combinations of words. An algorithm prioritizes the results and presents them to the translator according to what will provide the maximum reuse of previous translations while following rules about the relative quality of the source – certain document repositories can be rated above others, exact matches with approved terminology have priority over reference corpus matches, etc. This whole search process is one step and executes in less than a second – even when referencing sources that total millions of words. As always, the full original usage and style context of the matched expressions and their equivalent translations are provided. Reviewing and inserting the suggested translations is then a simple point-and-click operation.

This aggregated, automated search capability mines many sources of potentially valuable multilingual references, including:

- Previously translated original documents in WORD, HTML, PDF, or other formats;
- Multilingual web sites, such as the WHO trilingual site with thousands of words of high quality translations in the health field;
- Other sources of previously translated documents;
- Existing Translation Memory databases. In fact, because MultiTrans fully indexes these files too, users get all of the benefits of TM exact and fuzzy full sentence matches plus the ability to mine deeper by retrieving sub-sentence strings. MultiTrans also creates industry-standard format TMs as an output of the translation process.
- Terminology databases, project-specific lexicons and third-party glossaries.

7 Collaborative Language Management Infrastructure

All of the multi-language resources that MultiTrans puts at the fingertips of a translator with the workbench are also valuable to other professionals along the language management value chain. An n-tier and web-based environment enables, over a network or the Internet, several users to search, share, view and update the same corpus or terminology repository in real-time – the moment a user adds a term, that new language management asset can instantly be accessed and used by other language professionals –those with a MultiTrans workstation and those that are accessing in a read-only mode via the web.

Authors can reference the central language resources for monolingual assistance with term definition, style and usage and full-text examples of previous usage.

Teams of Translators distributed across in-house, agency and freelance organizations can share central multilingual resources.

Terminologists can leverage the resources for new terminology generation and research support during translation projects.

Editors and Reviewers can clarify terms, usage and see examples of previous translations.

Content Consumers who need to read and understand documentation can obtain clarification on terms and language by easily accessing definitions and examples.

8 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 been limited by their dependence on whole sentence repetition, their loss of translation context and the labor-intensity of verifying the initial TM databases.

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

The corpus-based approach enhances 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.

System Requirements - MultiTrans Workstation:
Windows 95, 98, Me, NT4.0 (SP4), 2000, XP
Pentium II 300 MHz
64 MB RAM
800x600 display
20MB free disk space
Internet Explorer 5.0 or higher

System Requirements - Client-Server:
Windows 2000 (SP2) or
Windows NT4.0 (SP4) with Windows NT4.0 Option Pack incl. IIS and MTS 2.0, and MDAC 2.5 (SP2)

System Requirements - Web:
MultiTrans TermBase and/or TransCorpora Server
IIS 5.0
ASP.net Framework