The SYSTRAN Linguistics Platform:

A Software Solution to Manage Multilingual Corporate Knowledge

White Paper

October 2002

I. Translation and Localization – New Challenges

Businesses are beginning to encounter multilingual information more frequently than in the past. Much of this information is related directly, or indirectly to the Internet and increasing corporate documentation. In some cases, businesses are reaching out to new markets, or are contacted by potential customers who do not speak their language. For other types of businesses, keeping up with business information in many languages is a competitive necessity. As a result, the diversity of translation and localization work is increasing. Software localization, customer support, intra-company communications, and customer communications are growing markets for translation products and services. The overall expenditure for worldwide localization, translation and interpretation services is projected to reach $5.8 billion by 2006, representing a CAGR of 14.6% from 2001 to 2006, according to IDC.

The changing nature of translation work has heightened interest in automated translation products and services. In the pre-Internet economy, most companies with translation requirements either outsourced their translation projects to external providers, or maintained in-house translation and localization staff. The latter option is costly, and is practiced only by large multinational companies. Both solutions are increasingly inadequate as the volume of information far outstrips the capacity for human translation.

SYSTRAN
Information and translation technologies

Headquarters
SYSTRAN S.A.
1, rue du Cimetière - BP7
95230 Soisy-sous-Montmorency
France
Tél. : + 33(0)1 39 34 97 97
Fax : + 33(0)1 39 89 49 34

North America
SYSTRAN Software Inc.
9333 Genesee Avenue
Plaza Level, Suite PL1
San Diego, CA 92121 - USA
Tel. : + 1 858 457 1900
Fax : + 1 858 457 0648

www.systransoft.com

The changing needs for business translation services novel solutions. Many new technical challenges have emerged, such as real-time delivery, colloquial text translation, speech translation and needs for uncommon language pairs.

Translation and localization projects for software and other products also involve complex linguistic and engineering challenges that may include text editing, format conversion, translation, cultural adaptation and script conversion. Very few businesses are staffed with the software engineers, linguists and translators needed to conduct a successful localization project. Costs aside, for information-based businesses, human translation is simply too slow and too costly to keep up with ever changing data. Time-sensitive applications such as news monitoring or stock trading can become obsolete in the hours, or even days that are needed for human translation. Multilingual customer support and communication also require on-demand translation, and often involve large volumes of text. Some examples are Internet self-help applications such as self-service technical support, email support and click-to-chat support. Users of these applications expect immediate results, a standard that would be impossible to meet with human translation. Even when translations are not needed immediately, it is prohibitively expensive to translate high-volume customer communications. Likewise, within companies, email and instant messages require rapid turnaround at high volumes. These cannot be addressed with human translation solutions.

Human translation and localization services are not feasible for highly dynamic content. Many software companies maintain large knowledge bases of support materials. This content changes substantially as new bug reports, fixes and upgrade information are added. The cost and delays of retranslating a large, dynamic content source is unsupportable for any business. Some companies have attempted to use Translation Memory (TM) systems in conjunction with content management software to manage translation of their content. This solution is not adequate for dynamic content.

SYSTRAN
Information and translation technologies

Headquarters
SYSTRAN S.A.
1, rue du Cimetière - BP7
95230 Soisy-sous-Montmorency
France
Tél. : + 33(0)1 39 34 97 97
Fax : + 33(0)1 39 89 49 34

North America
SYSTRAN Software Inc.
9333 Genesee Avenue
Plaza Level, Suite PL1
San Diego, CA 92121 - USA
Tel. : + 1 858 457 1900
Fax : + 1 858 457 0648

www.systransoft.com

TM systems rely on an existing resource of bilingual translation data. For example, a TM system can store a set of recurring sentences or phrases, and the translations that were made. Later these translations can be simply "plugged in", speeding the translation process.

For dynamic content, however, the percentage of reused content is low, making TM systems of limited value. Further, the TM system is not useful until it contains a large database of parallel source and target text. Producing the source and target text requires substantial effort by human translators. The cost and delay of human translation is unsupportable for any high volume or dynamic content. Automation of the translation process is the only viable solution for the majority of these applications. Yet many companies have resisted using machine translation because of concerns about quality, and the challenge of determining return on investment for an emerging technology. In some cases also, organized opposition by translators has been a factor in the decision not to use machine translation. The quality issue is central to the debate, and is the most frequently cited reason for not automating translation functions. SYSTRAN addresses this problem on multiple fronts, through integration support, custom lexical development, and quality review tools and metrics. SYSTRAN works with customers to ensure they understand the machine translation workflow. This is critical to ensuring a successful implementation because organizations that are new to MT often have difficulty foreseeing its impact.

II. The SYSTRAN Linguistics Platform

SYSTRAN's SLP (SYSTRAN Linguistics Platform) is a comprehensive enterprise solution for managing a full range of translation and localization project tasks. The SLP consists of the SYSTRAN machine translation (MT) technology, linguistic resources and tools for project management, corpus analysis and quality evaluation.

SYSTRAN
Information and translation technologies

Headquarters
SYSTRAN S.A.
1, rue du Cimetière - BP7
95230 Soisy-sous-Montmorency
France
Tél. : + 33(0)1 39 34 97 97
Fax : + 33(0)1 39 89 49 34

North America
SYSTRAN Software Inc.
9333 Genesee Avenue
Plaza Level, Suite PL1
San Diego, CA 92121 - USA
Tel. : + 1 858 457 1900
Fax : + 1 858 457 0648

www.systransoft.com

The underlying platform that supports the SLP is the SYSTRAN WebServer, a client/server application that can be accessed transparently through most common software applications. It supports document formats including HTML, RTF, XML, and SGML. The SYSTRAN WebServer is hosted at the customer's site and can be integrated with internal translation workflow systems. The SYSTRAN WebServer is a robust and high-volume platform that can support an unlimited number of users, and millions of translation jobs per day.

A. SYSTRAN MT System

The SYSTRAN machine translation system is a cornerstone component of the SLP. The system represents more than 30 years of accumulated knowledge and techniques for natural language processing. SYSTRAN has conducted a comprehensive rearchitecture of its system to leverage modern programming techniques and streamline access to its large linguistic and lexical resources.

SYSTRAN has the largest lexical resources and number of language pairs of any MT system in the world. Basic dictionaries contain up to 300,000 entries for each language. SYSTRAN also includes dozens of special subject dictionaries, many containing more than 100,000 entries. SYSTRAN's dictionaries are monolingual, a design feature that speeds creation of new language pairs because monolingual dictionaries can be recombined in new pairings.

The SYSTRAN MT system has received commercial acceptance for its high quality translation, robust architecture and broad range of language pairs and dictionaries. SYSTRAN has deployed its enterprise translation solution in high profile Internet sites such as AltaVista and Google and for corporate intranet applications at DaimlerChrysler and Dow Corning. The technology has also been used for decades by the U.S. Government and the European

Commission. These deployments have raised corporate, public and government awareness of machine translation technology, and have enabled millions of Web users to access multilingual Web content.

B. SLP Architecture and Properties

The SYSTRAN Linguistics Platform (SLP) is a fully customizable open platform for translation project management. The system's declarative, modular design organizes linguistic, engineering and project management functionality as a set of independent modules. As a result, users can craft custom solutions that use only the specific components that are required for an application. The SLP supports collaboration by allowing users to develop linguistic resources and manage and review translation projects throughout the translation lifecycle.

Because the SLP's resources are extremely large, a high degree of modularity is essential. Common office hardware cannot run the totality of the SYSTRAN resources, including its 36 language pairs and hundreds of dictionaries. For most applications, only a subset of the total resources is needed. The SLP's modularity allows users to install the components they need, utilizing typical hardware. In keeping with its modular design, the SLP supports the full XML standard, making possible the export of technology components for applications in content management, indexing and search and retrieval.

The SLP makes use of finite state technology to streamline computationally intensive tasks such as dictionary lookup and factorizing linguistic paradigms. This results in greater efficiency and performance for translation tasks.

Efficiency is also supported by the use of implicit transfer within the translation process. Implicit transfer addresses local expressions and verb phrases. These entities often are governed by a different set of syntactic

rules than what is generally applied to the source language. Implicit transfer sets up parallel source and target descriptions. During the transfer phase of translation, the descriptions are used to generate a syntactically correct structure in the target, rather than using a syntactic analysis of the internal structure of the expression.

Functional Components

The SLP architecture consists of two functional components, the Translation Processor, and the Translation Monitor.

SYSTRAN Translation Processor

The Translation Processor component provides the organizational framework for the SLP functions and resources, and the user's project files. Translation Processor is the core engine of the SLP. It consists of a set of scripts that support project creation, file management, corpus update and batch translation. The SLP is designed to handle large numbers of files. The SLP can produce reports on translation results, organized according to trends. This provides users with the ability to closely monitor and manage MT quality at levels ranging from gisting to high quality, publication-ready translation.

SYSTRAN Translation Monitor

The Translation Monitor component is a browser-based user interface to the SLP functionality. Using familiar Windows navigation conventions, users can access a full range of linguistic resources and project management tools, and direct the complete localization cycle for varied applications. SYSTRAN provides customization services tailored to the customer's content, delivery requirements and computing environment. With careful customization, SYSTRAN users can create high quality translations. The ability to produce high quality translation has been a key barrier to the entry of MT into many

Headquarters
SYSTRAN S.A.
1, rue du Cimetière - BP7
95230 Soisy-sous-Montmorency
France
Tél. : + 33(0)1 39 34 97 97
Fax : + 33(0)1 39 89 49 34

North America
SYSTRAN Software Inc.
9333 Genesee Avenue
Plaza Level, Suite PL1
San Diego, CA 92121 - USA
Tel. : + 1 858 457 1900
Fax : + 1 858 457 0648

www.systransoft.com

markets. The SYSTRAN Linguistics Platform removes this barrier, opening the doorway to a broad range of new MT applications.

Translation Monitor Features

The **Control Panel** functionality allows users to set the choice of target language for translation and launch batch translations. Batch translation is distinguished from interactive translation. In batch mode, the user supplies the system with the file name of the source text. The translation system retrieves and translates the file based on the user's configuration choices. Interactive translation, alternatively, performs a translation of text that is visible on screen, or passed to the translation system from an open application.

Once it is completed, users can view statistics on the translation job including the size of the text, processing time and listings of unfound words. These statistics provide important data for analyzing the success of the translation job. For example, if the percentage of unfound words is high, the user can enter the words in the system dictionary, then retranslate to create a better result. Understanding the areas of the text where the system had difficulty can also help to guide controlled writing efforts.

Quality evaluation tools are also accessible from the **Control Panel**. The **Control Panel** offers a baseline for comparing the quality of translations. As the user improves the system through addition of specialized vocabulary, improvements in quality can be tracked, ensuring that an acceptable result is produced.

The **Control Panel** contains tools for segmentation of large corpora into error categories. Categorization helps to make recurring patterns immediately obvious, so the reviewer can craft an encompassing solution.

Headquarters
SYSTRAN S.A.
1, rue du Cimetière - BP7
95230 Soisy-sous-Montmorency
France
Tél. : + 33(0)1 39 34 97 97
Fax : + 33(0)1 39 89 49 34

North America
SYSTRAN Software Inc.
9333 Genesee Avenue
Plaza Level, Suite PL1
San Diego, CA 92121 - USA
Tel. : + 1 858 457 1900
Fax : + 1 858 457 0648

www.systransoft.com

The **Analysis** tools available in the SLP support linguistic analysis of the user's corpora. Results of an analysis may be used to direct lexical development, guide controlled writing standards or indicate other SLP resources that are needed. For example, using one of the Analysis tools, the user can find and review multiple occurrences of a word within a corpus to understand its context and translation. Listings of words not found in the dictionaries can also be produced, along with testing of dictionary entries. Dictionary additions can have unintended impacts, making testing essential prior to including new terminology in permanent dictionaries.

The **Navigation** functions include a set of comparison tools for source and target text that allow the user to navigate through aligned corpora to find and compare translations of specific text segments.

The SLP's **Administration** capabilities include corpus definition, configuration of target language settings, dictionary selections and configuration of translation parameters.

Linguistic Resources

SYSTRAN's linguistic resources enable users to manage and customize the translation process from preparation of terminology, texts and dictionaries to post-MT processing. The linguistic components are agent-like, in that they perform specific types of linguistic tasks that can be initiated by the user. Agents can consult different modules, and even other agents in seeking to complete a task. As a multi-agent system, SYSTRAN can operate more efficiently than earlier MT system designs, which required explication of every possible interaction, and the triggering conditions. A high level of customizability results from this design, because the exact set of resources needed for a particular text type, language pair and communication requirement can be assembled.

SYSTRAN
Information and translation technologies

Headquarters
SYSTRAN S.A.
1, rue du Cimetière - BP7
95230 Soisy-sous-Montmorency
France
Tél. : + 33(0)1 39 34 97 97
Fax : + 33(0)1 39 89 49 34

North America
SYSTRAN Software Inc.
9333 Genesee Avenue
Plaza Level, Suite PL1
San Diego, CA 92121 - USA
Tel. : + 1 858 457 1900
Fax : + 1 858 457 0648

www.systransoft.com

Pre-Translation Resources

The **Document Filter** separates formatting codes from text prior to translation and reinserts them after the translation process. SYSTRAN's **pre-processor** identifies the document type, such as email, or structured vs. unstructured text. This information is useful for determining which additional resources are needed to achieve the best possible translation. For example, email and chat texts may have limited use of punctuation. Since punctuation provides important clues for segmentation, other means of segmenting sentences and sentential units would need to be employed. SYSTRAN supports numerous character encoding formats and performs conversion of character sets using its **character set converter**. A **language recognizer** can identify the language of the source text based on frequently occurring words. SYSTRAN also performs **spelling correction**. Spelling errors can have substantial negative impact on MT output quality. Misspelled words cannot be matched to entries in the dictionaries. As a result, these words go untranslated. A more insidious effect results from the fact that unknown words are assigned default part of speech. If the default part of speech is incorrect, other decisions concerning the syntactic structure of the sentence may be incorrect as well. The subject area of the text can be identified with the **semantic domain recognizer.** This tool consults SYSTRAN's extensive knowledge bases to identify terminology clusters that are characteristic of specific subject domains.

Translation Resources

Several natural language processing tools are used during the processes of sentence analysis, transfer and generation. The **sentence segmenter**

Information and translation technologies

Headquarters
SYSTRAN S.A.
1, rue du Cimetière - BP7
95230 Soisy-sous-Montmorency
France
Tél. : + 33(0)1 39 34 97 97
Fax : + 33(0)1 39 89 49 34

North America
SYSTRAN Software Inc.
9333 Genesee Avenue
Plaza Level, Suite PL1
San Diego, CA 92121 - USA
Tel. : + 1 858 457 1900
Fax : + 1 858 457 0648

www.systransoft.com

divides the text into sentences. This is a straightforward task if the text is composed exclusively of carefully punctuated, complete sentences. More commonly, sentence segmentation will have to account for missing punctuation, headings, tables of contents, greetings and closures and many other features of real-life text. After sentence segmentation, the **word delimiter** is used to determine word boundaries. For Western European languages, blank spaces are the primary indicator of a word boundary, though meaning units such as “ice cream” function as a single word despite their multi-word composition. However for many Asian languages, blank spaces are not used to mark word boundaries and word segmentation is a more complex task that utilizes morphological and syntactic information. Word forms are identified using a **lemmatizer** that matches the word form with morphological paradigms. To illustrate, the word form “known” would be identified as the past participle of the verb “know”. The **part-of-speech tagger** assigns a grammatical function (e.g., noun, verb, adjective etc) to each word based on its morphological features and the local environment of the word within the sentence. The **text synthesizer** performs the opposite function of the lemmatizer, generating correct word forms in the target language.

Post Translation Resources

Embedded within the SLP is the **SYSTRAN Review Manager** (**SRM**), a comprehensive toolset for evaluating translation quality and updating linguistic resources. The SRM organizes user’s corrections by category and performs statistical analysis on translation results. For example, a translated document can be analyzed on a sentence-by-sentence basis. Source and target languages are presented with color-coded markup to indicate not-found-words and the frequency of occurrence of a word or phrase. The reviewer can step through the translation sentence by sentence and provide feedback according to the nature and severity of the error. The SRM produces statistics based on user feedback and empirical characteristics of

the text, such as average sentence length, and word frequency. For errors that are based in dictionaries, a link is provided directly to the SYSTRAN Dictionary Manager. The reviewer can modify entries or make dictionary additions, then re-run the translation and review the results. After each review session, statistics are re-generated, allowing the reviewer to quantify translation quality, and the rate of improvement.

The QA tools include **terminology extraction**, **graphical editors** for development of syntactic and semantic rules, and examination of aligned corpora. **Benchmark tools** support users in creating benchmark standards and judging the quality of translations that are produced through the customization process, and during production.

Diverse Business Applications

The SLP has applications in numerous industries including customer support, knowledge management, cross language search and retrieval, multilingual enterprise applications, intra-company communications and eCommerce.

The cost of providing human technical customer support in many languages is becoming prohibitive. Automation with the SYSTRAN Linguistics Platform helps companies to achieve broader, more effective customer support at dramatically reduced costs.

SYSTRAN is well-versed in search and retrieval applications, having pioneered the application of MT to the U.S. Intelligence Community in 1968 and to AltaVista in 1998. SYSTRAN's technology is used by millions of Web users daily to translate foreign language websites on-the-fly. With the growing diversity of languages on the Web, translation of search results will inevitably become more widespread.

SYSTRAN
Information and translation technologies

Headquarters
SYSTRAN S.A.
1, rue du Cimetière - BP7
95230 Soisy-sous-Montmorency
France
Tél. : + 33(0)1 39 34 97 97
Fax : + 33(0)1 39 89 49 34

North America
SYSTRAN Software Inc.
9333 Genesee Avenue
Plaza Level, Suite PL1
San Diego, CA 92121 - USA
Tel. : + 1 858 457 1900
Fax : + 1 858 457 0648

www.systransoft.com

Within multinational businesses, SYSTRAN's technologies support intra-company communications such as MS Office Suite, LotusNotes, email, instant messaging (IM), short messaging services (SMS), chat, and message boards. In addition to person-to-person communications, corporate content such as human resources documentation, accounting data and internal manuals are excellent candidates for automated translation with the SLP. The cost and capacity of ongoing human translation of these materials is prohibitive, and deploying the SYSTRAN Linguistics Platform can produce a very rapid return on investment.

eCommerce applications have perhaps received more attention than any other potential MT application. Business-to-business as well as business-to-consumer applications can be launched successfully with the SLP. As an example, retailers with large catalogs of changing products can utilize the SLP to produce translations of their products at very low cost, reaching a broader audience.

The SYSTRAN Linguistics Platform opens the gateway for a host of new MT applications that are not possible using human translation or existing MT solutions. By demonstrating that high quality MT can be achieved, SYSTRAN has removed the chief obstacle to MT's success, paving a path for widespread business and consumer implementation of automated translation solutions.

SYSTRAN
Information and translation technologies

Headquarters
SYSTRAN S.A.
1, rue du Cimetière - BP7
95230 Soisy-sous-Montmorency
France
Tél. : + 33(0)1 39 34 97 97
Fax : + 33(0)1 39 89 49 34

North America
SYSTRAN Software Inc.
9333 Genesee Avenue
Plaza Level, Suite PL1
San Diego, CA 92121 - USA
Tel. : + 1 858 457 1900
Fax : + 1 858 457 0648

www.systransoft.com