



Database specifications:

- All text is stored in the UTF-8 version of the Unicode standard.
- The TMX (Translation Memory eXchange) standard has been adopted for database storage. TMX is an open XML standard developed in the commercial translation community to enable translators, translation service vendors, and clients to exchange and share translation resources.
- The Search and Match tools work with either MySQL or Oracle databases

Feature: TMX Format makes data portable!

Benefit: Databases created using the Language Weaver TM Generator can be imported into most commercial translation memory tools. Likewise, translation memories created using commercial tools can be saved in TMX format, and imported into the Language Weaver TM Generator.

Feature: Basis Technologies' RDIF (Rosette Document Ingestion Framework) provides automated format filtering, language recognition, and encoding recognition and conversion to UTF-8 Unicode.

Benefit: Text saved in your TM database is clean, and normalized to the Unicode standard which accommodates virtually all of the worlds written languages.

Hardware and OS Requirements The TM Generator Client and Server run on Microsoft Windows 2000/2003/XP. Minimum resource requirements for computers running the TM Generator are:

- 1.6 GHz processor
- 512 MB of RAM

Languages The TM Generator is language independent. Language ID, segmentation and tokenization capabilities are built in for: Arabic, Chinese, English, French, Hindi, Somali, and Spanish.

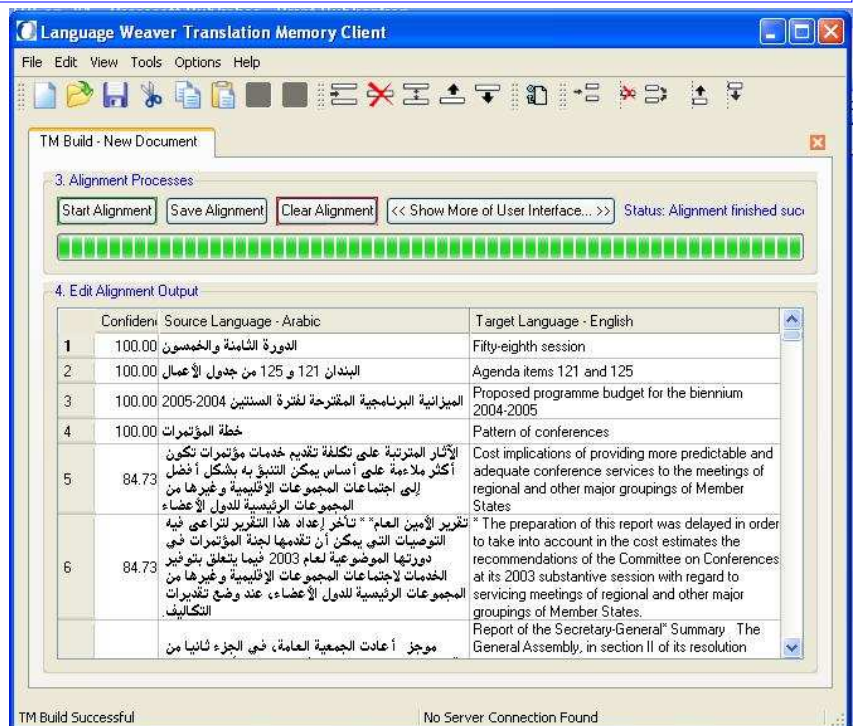
Language Weaver has developed special expertise in storing, processing, aligning and exploiting bilingual text resources—for example Arabic documents and their English translations. This expertise is a byproduct of our need for large quantities of sentence-aligned bilingual data to train our state of the art translation engines. However, many other groups also work with multilingual data and have accumulated translations. The TM Generator suite makes the corpus building and exploitation tools developed at Language Weaver available to the larger language community.

What is the TM Generator?

A suite of tools, built around a database application, for storing and reusing previously translated text. The bilingual corpora you create with the TM Generator can be exploited for reference—to see how others have solved translation problems in context; for productivity—to save and reuse frequently translated words, phrases and sentence; for standardization—to monitor how translators are translating important terms. The TM Generator suite currently provides three tools:

- **Build:** An interface and corpus alignment algorithm that allow users to align previously translated text. *Build* extracts text from common file formats (HTML, PDF, DOC, plaintext, XML—using Basis Technologies RDIF) and automatically aligns each text segment in English with its foreign language counterpart. Users may review and adjust alignments before saving the bilingual database as a TMX file. See the screenshot below. Words and phrases may also be saved.
- **Search:** Translations done by expert human translators provide an excellent opportunity for new translators, or translators working outside their subject area of expertise, to learn from the masters of their craft. *Search* gives linguists and learners access to the applied insights of other translators. (See reverse)
- **Match:** A productivity tool that utilizes previously translated material to semi-automatically translate new text. Somewhat like the commercial Translation Memory tools that gave the TM Generator its name, *Match* does not require full sentence repetitions in order to enhance productivity (See reverse.)

Build



Search leverages accumulated translations for reference. Users research the ways particular words or phrases have been translated before. This is particularly beneficial and appealing to human translators, and accelerates the return on investment in creating translation memory databases. Unlike conventional translation memory applications, which require nearly full sentence repetition before segments can be reused, every word that has been translated before has reference value.

Search supports multiple languages. A user can find the vernacular term for heroin in every language in the database. In this way, the search capability facilitates identification and standardization of important terminology. The results become electronic quick references lists.

Match enables users to reuse the translations of sentences and phrases that have been added to the TM database. The TM Match Engine will partially translate new documents by matching source sentences to segments saved in the TM database, and substitute translations from the TM database in the output. The output will be a mixture of foreign language and English sentences: Where matches were found in the TM database, the output will be the corresponding English segment. Where no match was found, the Arabic source text will be passed through untranslated.

Match 1: Where word or phrase segments are saved in the TM, they will be leveraged to partially translate sentences. Unmatched portions remain in the vernacular. Matched portions may require some editing for contextual agreement.

Match 2: The longest available match is selected, up to full sentences. The user can save the target English "Matched Text" for editing.

Search

Match 1

Match 2

