# The Temple Translator's Workstation Project*

Michelle Vanni
U.S. Department of Defense
mtvanni@afterlife.ncsc.mil

Rémi Zajac
Computing Research Laboratory
New Mexico State University
zajac@crl.nmsu.edu

The Temple project has developed an open multilingual architecture and software support for rapid development of extensible Machine Translation functionalities. The targeted languages are those for which Natural Language Processing and human resources are scarce or difficult to obtain. The goal is to support *rapid development of machine translation functionalities* in a very short time with limited resources.

The Temple Translator's Workstation is incorporated into a Tipster document management architecture and it allows both translator/analysts and monolingual analysts to use the machine-translation function for assessing the relevance of a translated document or otherwise using its information in the performance of other types of information processing. Translators can also use its output as a rough draft from which to begin the process of producing a translation, following up with specific post-editing functions.

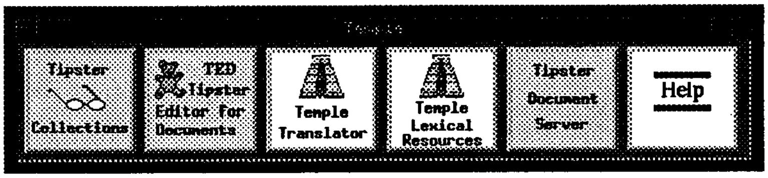


**Figure 1: The Temple tools.**

## Overview

Glossary-Based Machine-Translation (GBMT) was first developed at CMU as part of the Pangloss project [Nirenburg 95; Cohen et al., 93; Nirenburg et al., 93; Frederking et al., 93], and a sizeable Spanish-English GBMT system was implemented. The Temple project has built upon this experience and extended the GBMT approach to other languages: Japanese, Arabic, and Russian. This experience with other languages has provided significant insights for the development of a versatile GBMT engine and for the use of off-the-shelf components for building a complete Machine-Translation System. Building a generic platform for integrating various Machine-Translation Systems in a single flexible user environment built upon the Tipster document architecture [Grishman 95], has also been a valuable experience for developing generic Natural Language Processing support systems.

The user interface of the Temple Workstation includes a collection/document browser, the Tipster Editor for Documents, a generic translation function, access to lexical resources and context-sensitive help (Figure 1).

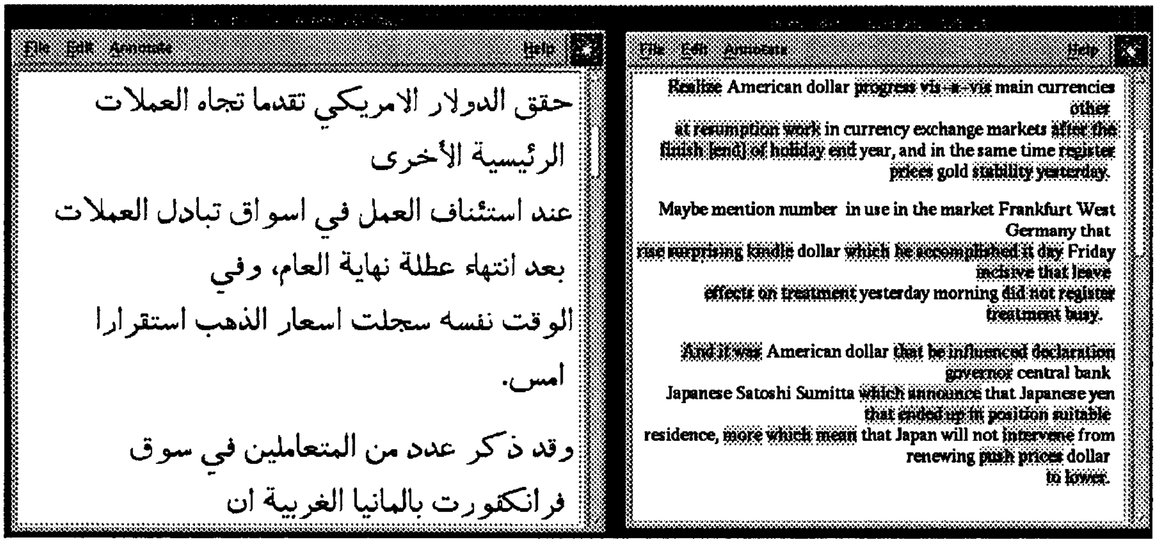
The Temple Translator's Workstation design is original in that it combines the best features and eliminates the weaknesses of competing alternatives. On the one hand, like word-based glossers, it puts the user in control by allowing all core linguistic components used by the glossary-based engine to be accessed, modified and developed by the translator. On the other hand, like advanced MT systems, it uses reliable morphological processors and taggers, components which are relatively inexpensive, require little or no maintenance, and greatly enhance output quality.

Currently, the Temple prototype provides automatic raw English translations from documents in several languages (Spanish, Arabic, Japanese and Russian). Translations are produced using a GBMT engine.

* Funded by DOD, Maryland Procurement Office, MDA904-94-R-3075/A0001.

Analysts and translators can edit the raw translation using a multilingual editor (Figure 2). Source documents and their translations are managed using the Tipster Document Manager developed at CRL, which is also used as the architectural basis for integrating the system's components.

The core components of the glossary-based engine are the bilingual dictionaries and the bilingual glossaries, which can easily include entries based on translators' own notes using a multilingual lexical database editor. It is this very database that is accessed at run time by the machine-translation system. Thus, when a translator modifies the lexical database (Figure 5), the modifications are immediately seen and used by the glossary-based engine in the machine-translation system. By contrast, in MAHT systems, dictionaries and glossaries are intended for human access only, and in almost all advanced MT systems, dictionaries (but not glossaries) can only be accessed and updated by a lexicologist with special training.



حقق الدولار الامريكي تقدما تجاه العملات
الرئيسية الأخرى
عند استئناف العمل في اسواق تبادل العملات
بعد انتهاء عطلة نهاية العام، وفي
الوقت نفسه سجلت اسعار الذهب استقرارا
امس.

وقد ذكر عدد من المتعاملين في سوق
فرانكفورت بالمانيا الغربية ان

Realise American dollar progress vis-a-vis main currencies
other
at resumption work in currency exchange markets after the
finish [end] of holiday end year, and in the same time register
prices gold stability yesterday.

Maybe mention number in use in the market Frankfurt West
Germany that
rise surprising kindle dollar which he accomplished it day Friday
incisive that leave
effects on treatment yesterday morning did not register
treatment busy.

And it was American dollar that he influenced declaration
governor central bank
Japanese Satoshi Sumitta which announce that Japanese yen
that ended up in position suitable
residence, more which mean that Japan will not intervene from
renewing push prices dollar
to lower.

**Figure 2: An Arabic document and its raw translation.**

## Architecture

The Temple prototype includes:

- A GBMT engine that provides an automatic translation for each language pair.
- Morphological analyzers, bilingual dictionaries, and bilingual glossaries for Spanish, Arabic, Japanese and Russian, and an English morphological generator [Penman 88].
- A multilingual document editor (the Tipster Editor for Documents developed at CRL under the Norm project) used to browse documents and their translation.
- A multilingual dictionary and glossary editor and utilities to parse and load flat dictionary (Machine-Readable Dictionaries) and glossary files into the system's lexical database.
- Corpus-based utilities to automatize the acquisition of bilingual glossaries.
- A Tipster Document Manager to support access and processing of user's documents.

The Temple architecture is capable of handling a large number of character codesets through the use of the multilingual text library developed at CRL, which includes a multilingual string library, a multilingual widget library (use for example to develop the multilingual lexical editor) and the multilingual Tipster Editor for Documents.

Tipster annotations are used as a lingua franca for representing linguistic information shared among various NLP components, such as morphological analyzers, taggers, bilingual dictionaries, the GBMT engine and the morphological generator. Each component has access to the common data structure through a unique interface provided by the Tipster Document Manager developed at CRL. NLP components are integrated in the architecture through TCL wrappers and filters that interface the component with the Temple representation stored as annotations in the Tipster Document Manager. Since most of the NLP components use linguistic representation that may widely differ, a single internal representation is used, e.g., for encoding part-of-speech, morphological features, etc. An NLP component interface to the document manager includes a mapping from the component representation to the Temple internal unique linguistic representation.

One important outcome of the Temple project is the development of an architecture to support the reuse of NLP tools and resources:

- Tools that are acquired from an external source, such as morphological analyzers, generators, or taggers, can be integrated in the system with a minimum of programming effort.
- Heterogeneous linguistic resources are parsed and mapped to a common multilingual representation.

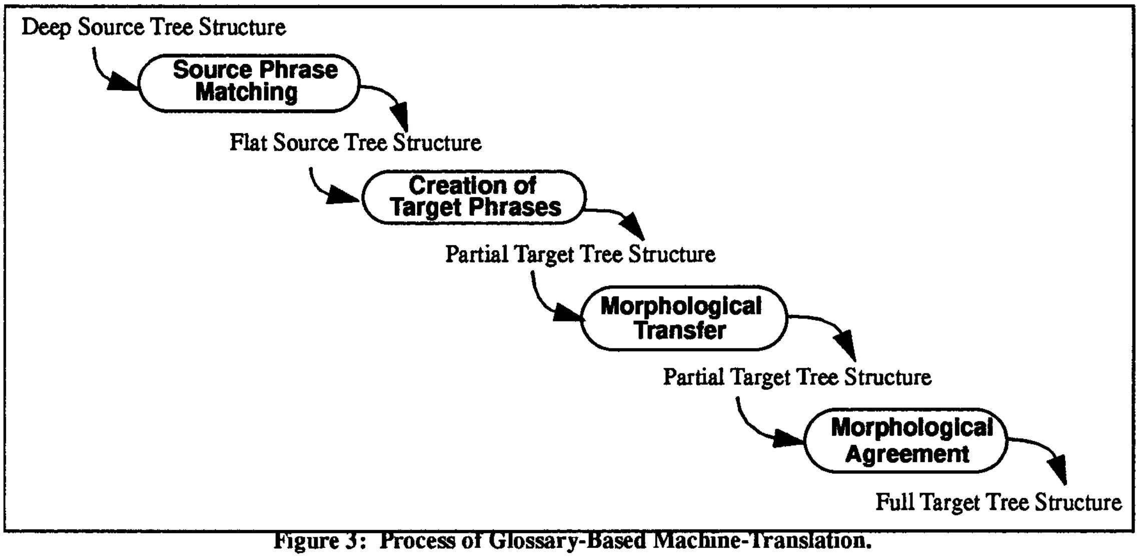

Deep Source Tree Structure
Source Phrase Matching
Flat Source Tree Structure
Creation of Target Phrases
Partial Target Tree Structure
Morphological Transfer
Partial Target Tree Structure
Morphological Agreement
Full Target Tree Structure

Figure 3: Process of Glossary-Based Machine-Translation.

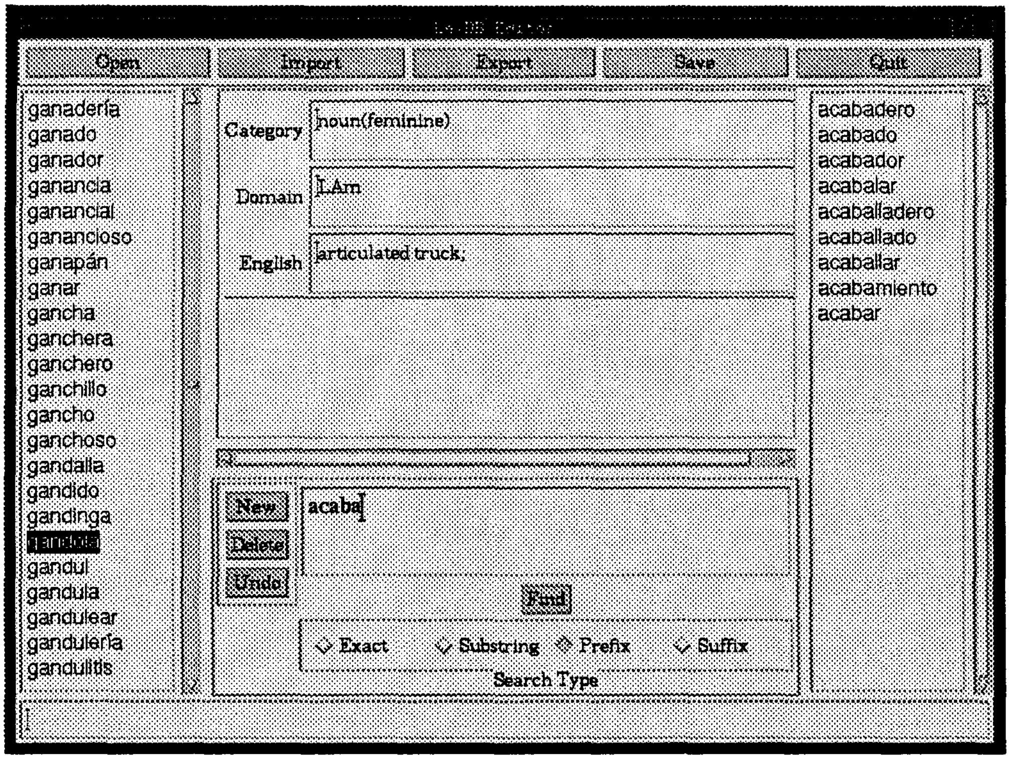
## Glossary-Based Machine Translation

The GBMT engine is the core component of the workstation machine-translation function. The GBMT engine is parametrized by a bilingual glossary. The bilingual glossary is essentially a phrasal dictionary: a glossary entry contains a source phrase pattern, a set of corresponding target phrase patterns, and correspondences between variables in the source and in the target patterns (Figure 4).

A GBMT system produces a phrase-by-phrase translation of the source text, falling back on a word-by-word translation when no phrase from the glossary matches the input. Thus, the size of the glossary and the flexibility of the pattern language are crucial for the production of better translations.The GBMT engine processes source tree structures in four steps:

1. Glossary phrases are matched within sentence sub-trees (produced by a morphological analyzer and various taggers and segmenters, depending on the language);
2. Target phrases patterns are added in the tree for each source phrase match;
3. Morphological information is transferred from source tokens to target tokens;
4. Agreement binding information is generated for each source phrase.

The tree structure manipulated by the GBMT engine contains both the source tree and the target tree which are simply source and target projections of the same data structure. Each target tree's lexical token is then sent to the morphological generator which produces the surface inflected form of each lexical token. Finally, the resulting fully instantiated tree structure is processed to produce a target Tipster document which contains alternative translations, tagging and morphological information and constituent information stored as Tipster annotations.



ganadería
ganado
ganador
ganancia
ganancial
ganancioso
ganapán
ganar
gancha
ganchera
ganchero
ganchillo
gancho
ganchoso
gandalla
gandido
gandinga
gandul
gandula
gandulear
gandulería
gandulitis

Category noun(feminine)

Domain LAm

English articulated truck;

acaba

◇ Exact ◇ Substring ◆ Prefix ◇ Suffix

Search Type

acabadero
acabado
acabador
acabalar
acaballadero
acaballado
acaballar
acabamiento
acabar

Figure 4: The lexical editor with the Spanish dictionary.

## Reuse of MRDs

Bilingual dictionaries that are used for the word-for-word fall-back translation are processed versions of various MRDs[1] (e.g. the Spanish-English Collins Dictionary, Figure 4) or of other MT dictionaries that have been restructured according to Temple own dictionary structure.

## Semi-automatic development of glossaries

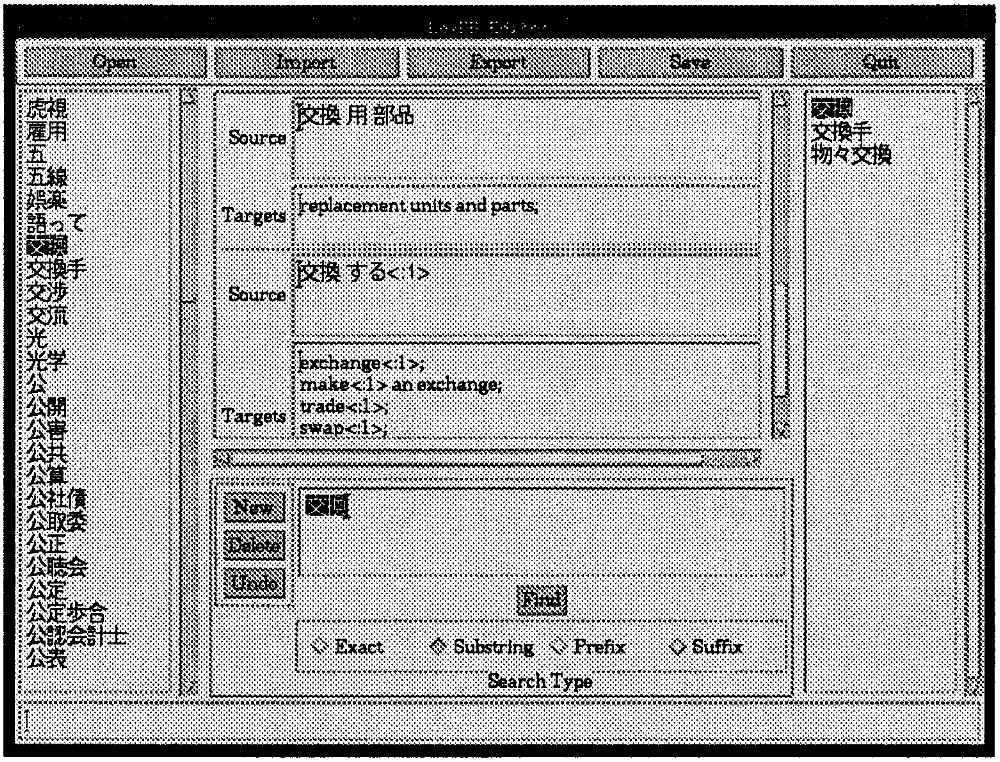
The availability of a large glossary is the key for good quality translations. The Temple Translator's Workstation provides the MT developer with tools to semi-automatically build glossaries. These tools work on large tagged corpora and use statistics on co-occurrence of words in a given corpus to extract phrase patterns.

The translator uses a phrase extraction utility to build a list of recurring patterns of words in a corpus (Ngrams). This list is formatted as a list of partial glossary entries and is then loaded in the lexical database. The translator can then use the glossary editor (Figure 5) to edit any entry flagged as incomplete. Using the glossary editor, the translator can also access bilingual dictionaries and use a variety of corpus-analysis tools, including a key word in context (KWIC) utility and a concordance tool.

The glossary is clearly dependent on the kind of text included in the corpus being used, but dependency on a particular domain and type of text is a natural limitation of machine-translation systems, and a GBMT is no exception. However, building a small size glossary, such as the Arabic-English glossary which contains approximately 10,000 entries, is a relatively easy and fast task. The Arabic-English glossary, for example, was built in six man/months. Moreover, it is fairly easy to enhance the glossary when new texts are being processed: these new texts can be added to the corpus and the corpus can be processed again to provide a new list of potential glossary entries. The

[1] See for example [Guthrie et al. 93a, Guthrie et al. 93b, Stein et al. 93].

translator can, of course, manually add any phrase to the glossary.



虎視
雇用
五
五線
娯楽
語って
交換
交換手
交渉
交流
光
光学
公
公開
公害
公共
公算
公社債
公取委
公正
公聴会
公定
公定歩合
公認会計士
公表

Source 交換用部品

Targets replacement units and parts;

Source 交換する<:1>

Targets exchange<:1>;
make<:1> an exchange;
trade<:1>;
swap<:1>;

交換

○ Exact ● Substring ○ Prefix ○ Suffix

交換
交換手
物々交換

Figure 5: The glossary editor with the Japanese glossary.

## Conclusion

The Temple Translator's Workstation has been developed in C within a two-year project at CRL. The project has provided valuable results and insights for the development of a flexible multilingual platform for Natural Language Processing. Bilingual dictionaries and glossaries have been developed for Spanish, Arabic, Japanese, and Russian. The project has produced a working multilingual Translator's Workstation prototype with complete machine translation functions for Spanish, Arabic, and Japanese to English, and some Russian morphological analysis. It has also resulted in the development of a language and tool integration methodology that facilitates the process of developing a new machine-translation system and integrating it in a translator's working environment. The translations produced answer the need for fast multilingual machine translation capabilities as required in information processing environments because the linguistic components of the system are derived from the very texts undergoing translation and analysis in the system.

## References

[Cohen et al. 93] Cohen, A., P. Cousseau, R. Frederking, D. Grannes, S. Khanna, C. McNeilly, S. Nirenburg, P. Shell and D. Waeltermann. "Translator's WorkStation User Document." Center for Machine Translation, Carnegie Mellon University, 1993.

[Frederking et al. 93] Frederking, R., D. Grannes, P. Cousseau, and S. Nirenburg. "An MAT Tool and Its Effectiveness." Proceedings of the DARPA Human Language Technology Workshop, Princeton, NJ, 1993.

[Grishman 95] Grishman, Ralph, editor. Tipster Phase II Architecture Design Document Version 1.52, July 1995. (http://cs.nyu.edu/tipster)

[Guthrie et al. 93a] Guthrie, Louise, Guthrie, Joe, Wilks, Yorick, Cowie, Jim, Farwell, David, Slator, Brian, and Bruce, Rebecca. "A research program on machine-tractable dictionaries and their application to text analysis." CRL Technical Report MCCS-92-249. 1993.

[Guthrie et al. 93b] Guthrie, Louise, Rauls, Venus, Luo, Tao, Bruce, Rebecca. "LEXI-CAD/CAM, A Tool for Lexicon Builders." CRL Technical Report MCCS-93-259. 1993.

[Nirenburg et al. 93] Nirenburg, S., P. Shell, A. Cohen, P. Cousseau, D. Grammes, C. McNeilly. "Multi-purpose Development and Operations Environments for Natural Language Applications." Proc. of the 3rd Conference on Applied Natural Language Processing (ANLP-93), Trento, Italy.

[Nirenburg 95] Nirenburg, Sergei, editor. "The PANGLOSS Mark III Machine-Translation System". CMU-CMT-95-145. A Joint Technical Report by NMSU CRL, USC ISI and CMU CMT. April 1995.

[Penman 88] The Penman Primer, User Guide, and Reference Manual. 1988. Unpublished USC/ISI documentation.

[Stein et al. 93] Stein, Gees C., Lin, Fang, Bruce, Rebecca, Weng, Fuliang, and Guthrie, Louise. "The Development of an Application Independent Lexicon: LexBase." CRL Technical Report MCCS-92-247. 1993.