Document generated on 04/19/2024 3:28 p.m.

Meta
Journal des traducteurs
Translators' Journal

META

Tools for Machine-Aided Translation: The CMU TWS

Sergei Nirenburg

Volume 37, Number 4, décembre 1992

Études et recherches en traductique / Studies and Researches in Machine Translation

URI: https://id.erudit.org/iderudit/003739ar
DOI: https://doi.org/10.7202/003739ar

See table of contents

Publisher(s)

Les Presses de l'Université de Montréal

ISSN

0026-0452 (print)
1492-1421 (digital)

Explore this journal

Cite this article

Nirenburg, S. (1992). Tools for Machine-Aided Translation: The CMU TWS. *Meta*, *37*(4), 709–720. https://doi.org/10.7202/003739ar

Article abstract

Interactive user environments have been a central efficiency-enhancing feature of many modern computer applications, including natural processing. There are two major kinds of people for whom environments can and must be constructed — developers and end users. Developers need help with a variety of knowledge acquisition tasks, such as dictionary and grammar writing. End users (e.g., technical writers or translators) look for efficiency enhancements beyond the level of word processing support. There are two approaches to building interactive tools. A dedicated workstation can be developed for each of the required functionalities. Alternatively, workstations can be configured as sets of application routines attached to a universal user interface. In this paper, we describe a general-purpose user environment, under development at the Center for Machine Translation of Carnegie Mellon University. This environment can support several machine-aided human translation configurations and a human-aided machine translation configuration. It also supports knowledge acquisition for machine translation and other NLP systems.

Tous droits réservés © Les Presses de l'Université de Montréal, 1992

This document is protected by copyright law. Use of the services of Érudit (including reproduction) is subject to its terms and conditions, which can be viewed online.

https://apropos.erudit.org/en/users/policy-on-use/

érudit

This article is disseminated and preserved by Érudit.

Érudit is a non-profit inter-university consortium of the Université de Montréal, Université Laval, and the Université du Québec à Montréal. Its mission is to promote and disseminate research.

https://www.erudit.org/en/

TOOLS FOR MACHINE-AIDED TRANSLATION: THE CMU TWS

SERGEI NIRENBURG
Carnegie Mellon University, Pittsburgh, USA

Résumé

Les environnements permettant l'interaction avec l'utilisateur sont devenus un élément majeur dans l'amélioration de l'efficacité de nombreuses applications informatiques, y compris pour le traitement automatique du langage naturel. On trouve deux principaux types de personnes pour qui des environnements peuvent et doivent être construits — les développeurs et les utilisateurs. Les développeurs ont besoin d'aide pour une variété de tâches concernant l'acquisition des connaissances, telles que l'élaboration de dictionnaires et de grammaires. Les utilisateurs (par exemple, les rédacteurs techniques et les traducteurs) recherchent des aides pour l'amélioration de l'efficacité qui aillent au-delà du simple traitement de texte. Il y a deux approches à la construction d'outils interactifs. Une station de travail peut être développée spécialement pour chacune des fonctionalités requises. Alternativement, des stations de travail peuvent être configurées en ensembles de routines d'application reliées à une interface universelle.

Dans cet article, nous décrivons un environnement à but général, développé au Centre de traduction automatique de l'Université Carnegie Mellon. Cet environnement peut soutenir plusieurs configurations de traduction humaine assistée par l'homme. Il permet aussi l'acquisition de connaissances pour la traduction automatique et d'autres systèmes de TAL.

Abstract

Interactive user environments have been a central efficiency-enhancing feature of many modern computer applications, including natural processing. There are two major kinds of people for whom environments can and must be constructed — developers and end users. Developers need help with a variety of knowledge acquisition tasks, such as dictionary and grammar writing. End users (e.g., technical writers or translators) look for efficiency enhancements beyond the level of word processing support. There are two approaches to building interactive tools. A dedicated workstation can be developed for each of the required functionalities. Alternatively, workstations can be configured as sets of application routines attached to a universal user interface.

In this paper, we describe a general-purpose user environment, under development at the Center for Machine Translation of Carnegie Mellon University. This environment can support several machine-aided human translation configurations and a human-aided machine translation configuration. It also supports knowledge acquisition for machine translation and other NLP systems.

1. GENERAL

Interactive user environments have been a central efficiency-enhancing feature in the development of many modern computer applications, including natural language processing. At the same time, such environments can be seen as tools for enhancing the productivity of technical writers, translators, editors, lexicographers and other document production personnel.

Developers need help with a variety of knowledge acquisition tasks, such as dictionary and grammar writing. End users look for efficiency enhancements beyond the current level of word processing support.

There are two approaches to building interactive NLP tools. A dedicated workstation can be developed for each of the required functionalities. Alternatively, workstations can be configured as sets of application routines attached to a universal user interface. The Translation Workstation (TWS) project at CMU has opted for the latter choice.[1] In what follows we describe the tool functionalities that have been developed and some workstation configurations in which they will be used.

TWS consists of a(n ever growing) number of application (functionality) modules which are integrated through the central user interface module (see Nirenburg *et al.* 1992 for a more technical description). In what follows we introduce some of TWS functionalities. The description is organized around the set of intended TWS configurations — the "base model," an enhanced machine-aided human translation environment, a human-aided machine translation environment and a set of tools for MT system developers.[2]

A different approach to translator's workstation development has been taken by researchers at the Canadian Workplace Automation Research Center (*e.g.* Macklovitch 1989), where it was decided to configure a system from commercially available components. While development costs can be somewhat lowered with this approach, it is difficult to attain the level of interpenetration of the various modules and thus further limitations are imposed on the possibilities of efficiency enhancements. Other work on a similar package has been reported (*e.g.* Kugler *et al.* 1992).

2. TWS CONFIGURATION I: BASIC FUNCTIONALITIES

The set of basic functionalities of TWS includes customizable text editing facilities, access to online reference sources (dictionaries, text archives, encyclopedias, parts catalogs, etc.), a facility for graphically acquiring, extending and maintaining proprietory glossaries and a package for manipulating large text archives (essentially, frequency counts a "key word in context" utility, etc.) Figure 1 shows a sample layout of three TWS windows: The TEXT EDITOR, the DICTIONARY INTERFACE, and the WORD FREQUENCY windows. Each of the corresponding modules is described in additional detail below.

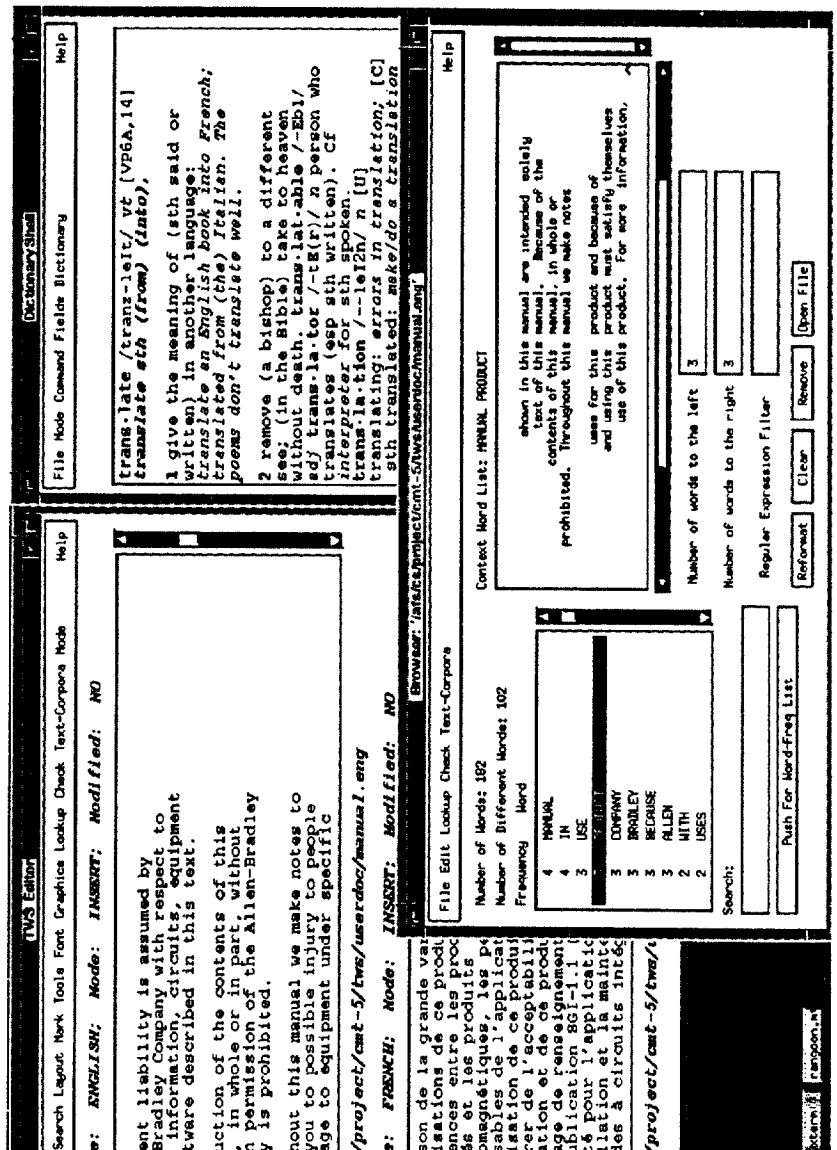

Figure 1:
Sample TWS Windows. The text-editing, dictionary and open-text modules each use their own window.

2.1. THE TEXT EDITOR

The editor facility is one of the most basic functionalities for both end-user and developer workstation configurations. The TWS editor's central feature is the capability to emulate various existing word processors. This feature is important from the standpoint of eventual acceptance of the workstation environments by end users. Currently supported word processor emulation modes include WordPerfect™, VI and Emacs. The emulation modes were reverse-engineered. That is, no part of the design or actual code of the word processors that TWS emulates was used in TWS itself. Synchronized scrolling in several editor windows is a feature that helps both translators and editors of translations.

2.2. THE DICTIONARY INTERFACE

TWS supports access to read-only materials (*e.g.* DICTIONARIES) and reference files which the user can modify (*e.g.* USER GLOSSARIES).

The DICTIONARY functionality provides the ability to access and display entries from machine readable dictionaries (both mono- and multilingual). After a word is selected, either by highlighting it in the editor window or typing it in a special dialog box, a dictionary window will appear with the definition of the word displayed (Figure 2).

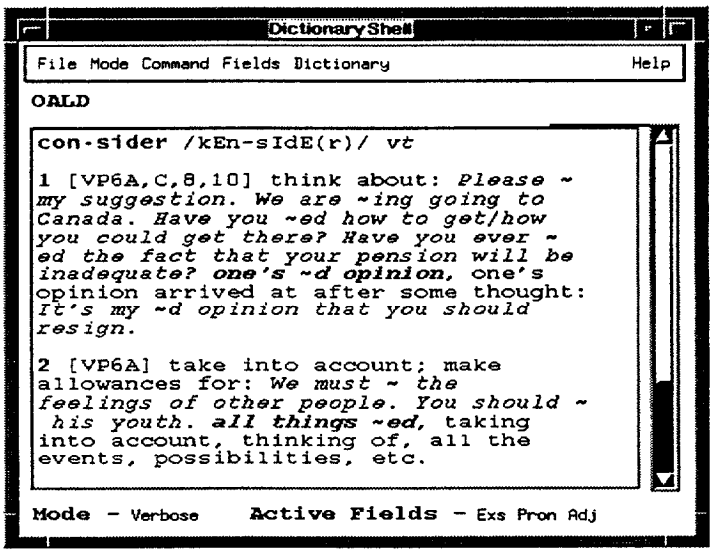

Figure 2:
The TWS dictionary interface showing the entry for *consider* in the Oxford Advanced Learner's Dictionary of English.

Three display modes (*terse, standard* and *verbose*) are implemented. The TWS includes, as a back-end functionality, a parser for an SGML-marked text which produces the necessary indices for flexible storage and retrieval of entries and entry parts. As SGML conventions allow for variability, introduction of new SGML-compatible reference materials will involve modifications to the parser.

The USER GLOSSARY functionality allows the user to create, modify, and maintain any number of personal reference materials. In the translation environment, we expect that most of these will be glossaries containing definitions and/or translations for words,

phrases, and/or idioms in the source text. One can retrieve a word or phrase from a glossary by highlighting it in the text editor and choosing an appropriate glossary in the menu.

The glossary is organized as an alphabetized list of entries. Each entry contains the following information:

- an English term;
- the corresponding terms in each of the required languages;
- a set of usage examples of all of the above terms (optional); and
- an explanatory description of the concept to which the term refers, in English (optional).

Once accessed, the glossary will be displayed in a single-pane editor window, with the cursor positioned at the word or phrase with which it was called. To add a new term to the glossary, the user chooses the *Add a Term* menu option and the entered term will appear in the alphabetically correct position in the file and the cursor will be placed immediately after it, ready for the user to type in the translation or explanation.

2.3. OPEN TEXT PROCESSING

This functionality is at present quite simple. It produces frequency counts and keyword-in-context (KWIC) listings. For a practical translator, this application is helpful for retrieval, from an archive of past translations, of a section similar to the one which must be translated currently. As it is possible to get to the appropriate place in the file from a line in the KWIC window, complete paragraphs and even larger sections of text could be moved into the current translation. This facility has a powerful keystroke-saving potential. In a development environment, this facility is useful for corpus analysis and vocabulary determination prior to lexicon acquisition.

3. TWS CONFIGURATION II: MACHINE-AIDED HUMAN TRANSLATION (MAHT)

In developing tools for end users one must take into account an entire production process. In the translation business, the process often starts before a document is submitted for translation. A lion's share of material translated around the world is documentation accompanying various products — typically, operator's and repair manuals for various types of equipment. The manuals have to be written and then updated and revised for every new model or version of the equipment. This offers opportunities for efficiency enhancements.

One of the most important tasks in a multilingual document production environment is to minimize the effort required for introducing revisions into existing documents. If the revised documents are intended for translation, it is desirable not to have those portions which are identical to an old version of the document translated again. Instead, old translations could be used. Identity of text components can be determined automatically. This functionality is a component of TWS-MAHT. This functionality is, however, much easier to understand than to implement, as differences among texts come in many varieties. Paragraphs, sentences and words can be added, deleted or changed; layout can be changed; graphics and tables can be added, deleted or changed, paragraphs and entire pages can be moved to a different place in a document, etc.

TWS-MAHT can be useful not only to revisors but also to technical writers and translators. A major interactive efficiency-enhancing facility for a *technical writer* is the ability to retrieve from an online archive paragraphs, sentences and phrases which include a given word or phrase. This necessitates a set of algorithms for search and retrieval of information from the archive. A technical writer will also benefit from a grammar and style filter which will automatically flag sentences and phrases which include nonstan-

dard technical language or convoluted phrase structure (TWS does not at present include this functionality).

When configured for a *translator*, TWS-MAHT includes all the above functionalities. In addition, the text comparison module is somewhat modified, so that it supports a version of the so-called "example-based" translation (several experiments with this approach have been reported — *e.g.* Brown *et al.* 1991, Sato and Nagao 1990, Kugler *et al.* 1991). In particular, in this configuration there is no presumption that a source text is a new version of another known document. An archive of previously translated materials is of central importance as the knowledge source for this functionality. Figures 3 through 7 illustrate the process of operation of the example-based translation module. This module is presently under development.

A 4-position keyswitch on the front panel of the processor module (Figure 2.1) lets you select one of four modes of operations

Figure 3:
A paragraph for translation.

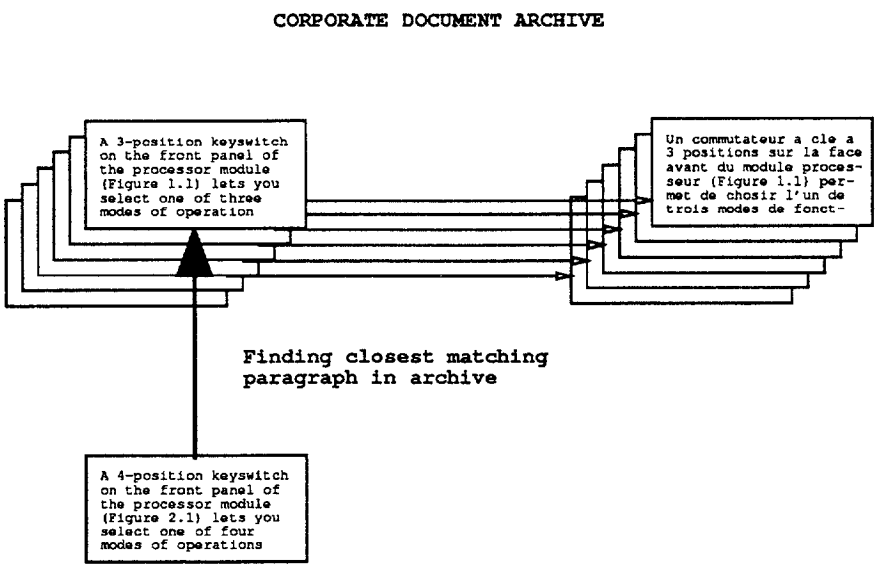

Figure 4:
Finding Closest Match.

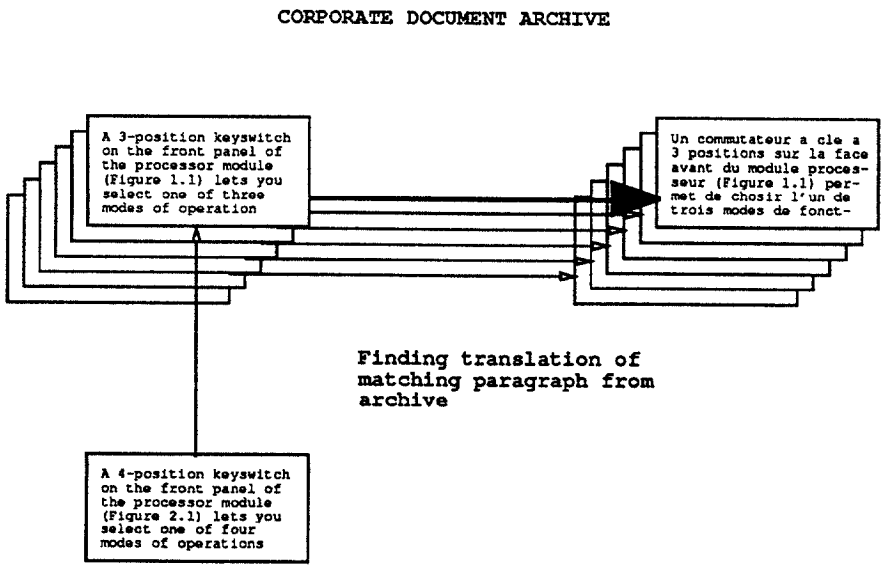

Figure 5:
Retrieving Translation of closest Match.

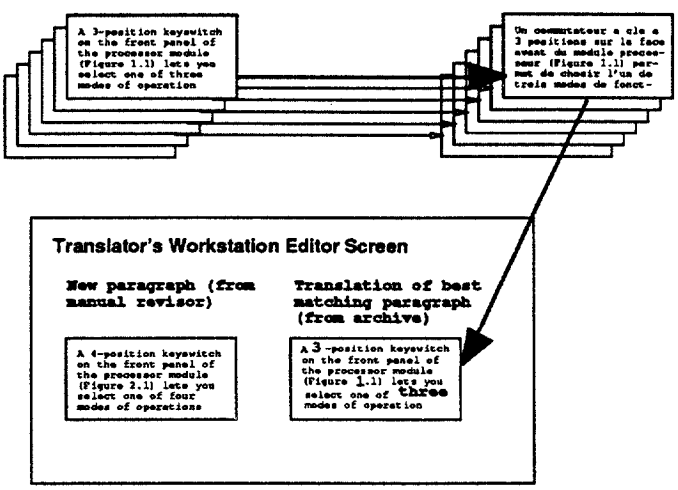

Figure 6:
Retrieving Translation of Closest Match into Editor.

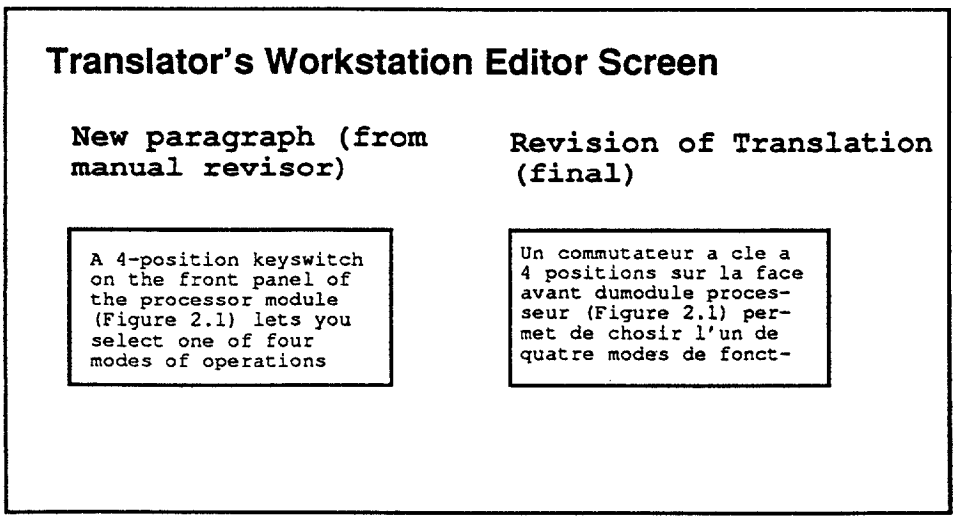

Figure 7:
Final Result of Editing Old Translation.

4. TWS CONFIGURATION III: HUMAN-AIDED MACHINE TRANSLATION (HAMT)

The MAHT configuration of TWS can support MT postediting. Text translated by machine can be displayed in the target text window, and the posteditor can work on it using the same tools as a translator would. Our approach to human-aided machine translation (HAMT) is different[3]. Unlike most other approaches, it does not rely on postediting machine output and therefore requires a different TWS configuration. In our approach, while a computer tries to translate a document, a human translator monitors its progress and gives the system guidance when called upon to do so. The primary role of the human is a) to help the system make processing choices for which it lacks sufficient knowledge and b) to update the lexicons to cover unexpected vocabulary.

The human-computer interface that supports a subset of such functionalities has been implemented on a small scale in the MIND machine translation project at the RAND Corporation (Kay 1973; the module was called "disambiguator"). A device of this sort was also discussed by Tomita (1986). A slightly larger interactive module was implemented in the KBMT-89 MT system, where it was known as the "augmentor" (Brown 1991).

Figure 8 illustrates the way TWS-HAMT carries out human-computer interaction in this system configuration. The dialog box with the disambiguation question is produced by the system and the user is supposed to click the appropriate box for the correct interlingua text to be produced. After the automatic synthesis stage, the translation will appear, sentence by sentence, in the right-hand side editor window, for the user to check.

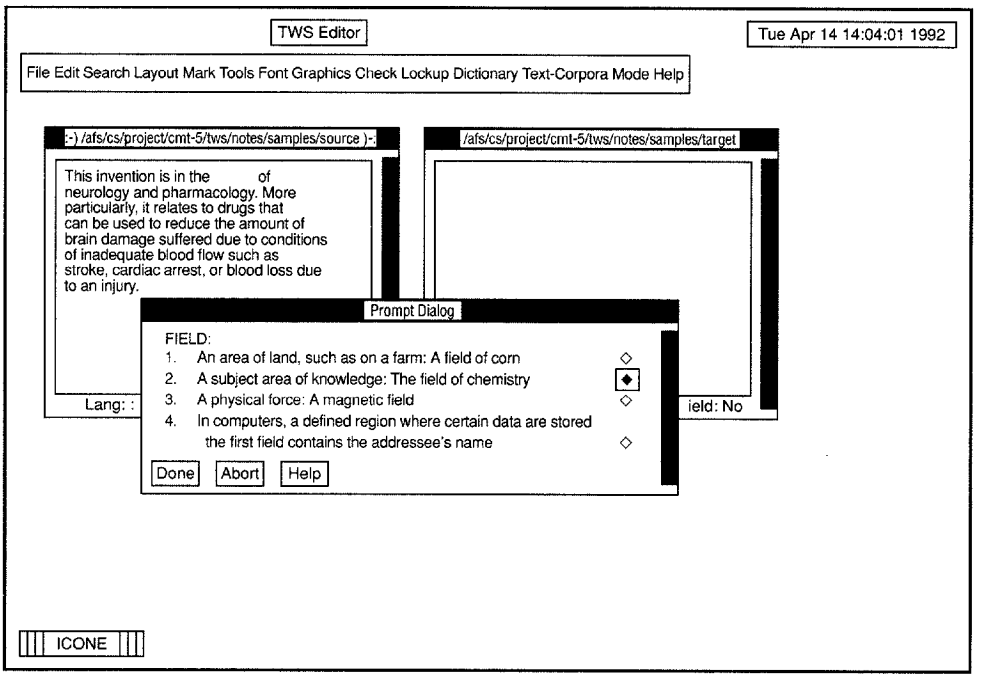

Figure 8:
User Interaction in Human-Aided Machine Translation.

5. KNOWLEDGE ACQUISITION SUPPORT

TWS is useful not only as an end user tool but also as a tool that enhances the efficiency of knowledge acquisition for machine translation (and, more generally, natural language processing) systems. A number of interactive NLP development environments have been reported (*e.g.* IRACQ (Ayuso *et al.* 1987), LUKE (Wroblewski and Rich 1988) or ONTOS (Monarch 1989)). All of them are meant to support a particular kind of NLP system. Thus, they trade generality and adaptability to new tasks for efficiency in solving the tasks of a given project. TWS provides a convenient graphics interface which can be used with more than one underlying system. In what follows we illustrate several of the knowledge acquisition environments which TWS can support.

5.1. ACQUIRING TEXTUAL DATABASES

The centerpiece of our approach to machine-aided human translation is the systems's capability to suggest possible translations for fragments of input text which the operator can use, modify or reject.

The capability of producing suggested best-estimate translations for human judgment is predicated on the availability of a large number of previously translated texts in relevant sublanguages. Once such texts are obtained, they are organized in an internal archive which is used for a variety of search and match operations resulting in the display of the "best fit" translated sentence for each sentence of input. The structure of this archive is illustrated in Figure 9. The management system for this database is currently under development.

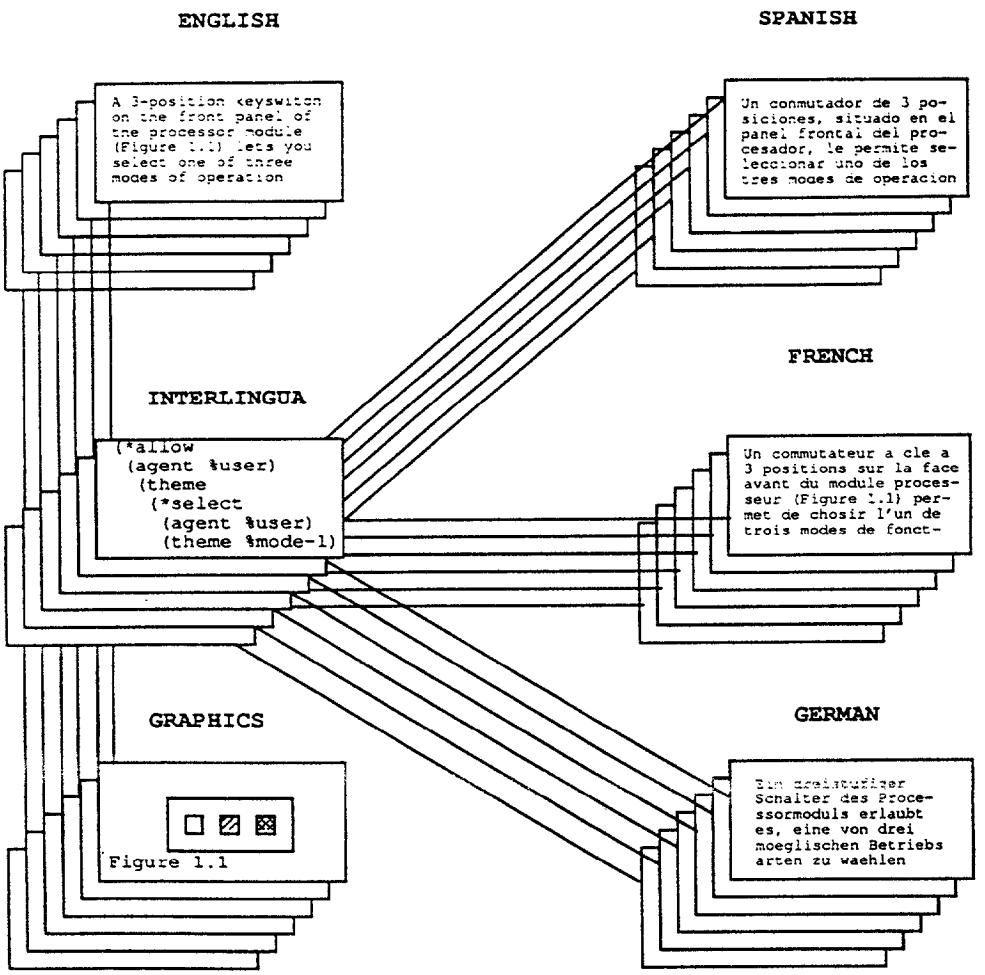

Figure 9:
The Structure of a Corporate Document Archive.

5.2. ACQUIRING ONTOLOGY: ONTOS

A central knowledge source in knowledge-based MT systems is a model of the world, often called "ontology". Ontology acquisition is a very time-consuming process. An ontology acquisition tool, ONTOS, has been developed at CMU (see, *e.g.* Monarch and Nirenburg 1988). ONTOS consists of a grapher, a knowledge manipulation system supporting the representation of multiple-inheritance hierarchies and an automatic checker for testing the correctness of the syntax and semantics of the knowledge elements.

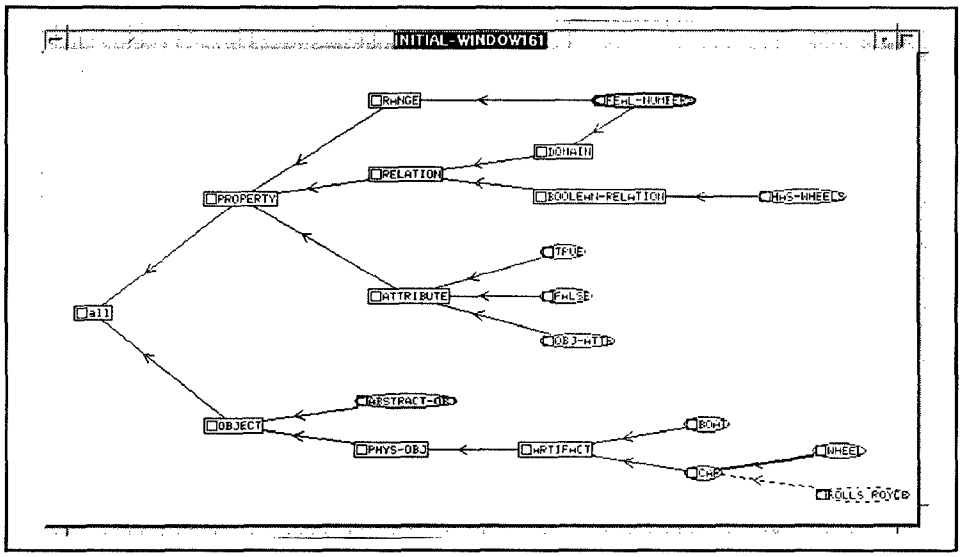

Figure 10:
A sample ontology.

Various browsing and acquisition actions can be performed by selecting objects, and then clicking on a menu item or pressing a key. The actions include creating, deleting, copying and updating nodes and links. The user can also edit the representations of conceptual objects corresponding to the various nodes and links.

Figure 10 illustrates the grapher interface. It displays a simple ONTOS ontology. ONTOS has been used to create a number of different ontologies for several projects, including the machine translation project KBMT-89 (see Goodman and Nirenburg 1992). A description of our approach to ontology building can be found in Carlson and Nirenburg, 1991.

5.3. LEXICON ACQUISITION

Acquisition of machine-tractable lexicons is as important for machine translation development as acquisition of ontologies. Within TWS, we have developed a module called LAI, the Lexicon Acquisition Interface to support this task. LAI consists essentially of a set of user questionnaires organized in decision trees and presented to the user through TWS dialog boxes and an entry browser window. The questionnaires "walk" the user through the list of morphological, syntactic, semantic and other properties required for a particular lexicon entry format. The system records the information obtained from the user in a specified formalism and displays the nascent lexicon entry in a browser window. This type of functionality has been implemented and used in a number of NLP systems. See Knight 1991 for a survey.

At present, our LAI supports the acquisition of lexicons only for English and for one particular lexicon format. In order to facilitate the creation of LAI questionnaires and browser windows for other languages and other kinds of lexicons, a special interface-building interface (IBI) has been implemented, which allows for interactive specification of menus and window widget sets (such as the dialog boxes and the browser window).

6. CONCLUSION

Translator workstations are at present in their infancy, but there is hope that in five to ten years their general availability on popular hardware platforms will forever change the nature of the translation industry.

A user environment, such as a TWS, can never be considered completed. There are always new and exciting functionalities that can be added to any system to make it smarter, faster, more flexible, customizable or convenient in use. In the CMU TWS project we have a long list of improvements to the system along the above lines. To date, the TWS has undergone only in-house testing. We plan to have it exposed to extensive practical tests by actual translators since only this type of feedback can give us real insights into development priorities.

Notes

1. The name TWS is, of course, quite restrictive, as our workstations are designed to help a broad spectrum of users, not just translators.
2. The TWS project is, naturally, a team effort. The TWS development team has included, at various times, Ariel Cohen, Peter Cousseau, Bob Frederking, Dean Grannes, Chris McNeilly and Pete Shell. Many thanks to all of the above for help in preparing this report. All the remaining errors and inconsistencies are, naturally, mine.
3. See Nirenburg *et al.* 1992 for a detailed description of our approach to translation, called Knowledge-Based Machine Translation.

BIBLIOGRAPHY

AYUSO, D., SHAKED, V. and R. WEISCHEDEL (1987): "An Environment for Acquiring Semantic Information", *Proceedings of Annual Meeting of ACL*, Stanford, CA.

BROWN, R. (1991): "Augmentation", K. Goodman and S. Nirenburg (Eds.), *KBMT-89: A Case Study in Knowledge-Based Machine Translation*, San Mateo, CA, Morgan Kaufmann.

BROWN, P., COCKE, J., DELLA PIETRA, J., DELLA PIETRA, S., JELINEK, F., LAFFERTY, J., MERCER, R. and P. ROOSSIN (1990): "A Statistical Approach to Machine Translation", *Computational Linguistics*, 16, pp. 79-85.

CARLSON, L. and S. NIRENBURG (1990): *World Modeling for NLP. Technical Report CMU-CMT-90-121*, Center for Machine Translation, Carnegie Mellon University.

KAY, M. (1973): "The MIND System", R. Rustin (Ed.), *Natural Language Processing*, New York, Algorithmics Press, pp. 155-188.

KNIGHT, K. (1991): *Integrating Knowledge Acquisition and Language Acquisition*, CMU PhD Thesis.

KUGLER, M., HEYER, G., KESE, R., VON KLEIST-RETZOW, B. and G. WINKELMANN (1992): "The Translator's Workbench: An Environment for Multi-Lingual Text Processing and Translation", S. Nirenburg (Ed.), *Progress in Machine Translation*, Amsterdam, IOS Publications, pp. 185-189.

MACKLOVITCH, E. (1989): "An Off-the-Shelf Workstation for Translators", *Proceedings of the 30th American Translators Conference*, Washington, D.C.

MONARCH, I. (1989): *ONTOS: Reference Manual. Technical Report*, Center for Machine Translation, Carnegie Mellon University.

NIRENBURG, S. and C. DEFRISE (1992): "Application-Oriented Computational Semantics, to appear in Johnson, R. and Rosner M. (Eds.), *Computational Linguistics and Formal Semantics*, Cambridge, Cambridge University Press.

SATO, S. and M. NAGAO (1990): "Toward Memory-Based Translation", *Proceedings of Coling-90*, Helsinki.

TOMITA, M. (1986): "Sentence Disambiguation by Asking", *Computers and Translation*, 1, pp. 39-51.

WROBLEWSKI, D. A. and E. A. RICH (1988): "Luke: An Experiment in the Early Integration of Natural Language Processing", *Proceedings of Second Conference on Applied Natural Language Processing*, Austin, Texas.