Oleada: User-Centered TIPSTER Technology for Language Instruction

William C. Ogden and Philip Bernick
The Computing Research Laboratory at New Mexico State University
Box 30001, Department 3CRL, Las Cruces, New Mexico 88003
email: ogden | pbernick@crl.nmsu.edu
phone: 505.646.5466

1.0 Abstract

TIPSTER is an ARPA sponsored program that seeks to develop methods and tools that support analysts in their efforts to filter, process, and analyze ever increasing quantities of text-based information. To this end, government sponsors, contractors, and developers are working to design an architecture specification that makes it possible for natural language processing techniques and tools, from a variety sources, to be integrated, shared, and configured by end-users. The Computing Research Laboratory (CRL) is a longtime contributor to TIPSTER. A significant portion of CRL's research involves work on a variety of natural language processing problems, human-computer interaction, and problems associated with getting technology into the hands of end-users. CRL is using TIPSTER technology to develop OLEADA, which is an integrated set of computer tools designed to support language learners, and instructors. Further, OLEADA has been developed using a task-oriented user-centered design methodology. This paper describes the methodology used to develop OLEADA and the current system's capabilities.

2.0 TIPSTER and the Computing Research Laboratory

Information extraction is a relatively new application of natural language processing techniques in which basic information and relationships are found and extracted from text. TIPSTER I was an effort to find electronic methods for information retrieval and information extraction.TIPSTER uses texts from a variety of sources including newspaper articles and wire service reports. The information TIPSTER I extracts resembles a completed form. The contents of a form is intended to be used to automatically generate specialized databases for information analysts. The components developed for TIPSTER I enabled it to function in two languages (Japanese and English).

TIPSTER II is a joint effort among many sites to develop working systems that integrate information retrieval and information extraction. The core of the project is a joint government/contractor committee whose goal is to specify an architecture for TIPSTER II. TIPSTER developers work to provide a variety of specialized software subsystems that support TIPSTER development. These include:

- Document managers that provide multi-source document compatibilities.
- Translation subsystems that support retrieval of documents in many languages, based on a query in one language.
- Libraries of procedures for user interface support with embedded functionality for Information Retrieval and Information Extraction.
- Advanced Motif-based multi-lingual user interface capabilities, supporting Chinese, Japanese, Korean, Arabic and other writing systems.
- Plug-and-play to integrate various kinds of software inside TIPSTER.

The TIPSTER II architecture makes it possible to integrate a variety of information retrieval, extraction, and text processing systems in ways that help analysts address more complex problems.

2.1 OLEADA: Task-Oriented User-Centered Design in Natural Language Processing

Researchers with the Computing Research Laboratory (CRL) at New Mexico State University are interested not only in theoretical aspects of natural language processing, but methods for getting the results of this research into the hands of actual users. OLEADA, a project at CRL that seeks to develop computer tools that support language learners and instructors, has been developed with this goal in mind. OLEADA uses TIPSTER technology to provide users with access to pertinent and authentic text, and tools for manipulating this text. It is based, in part, on Cibola, a system that supports human translators by providing them with tools that directly support the translation task. Although task-oriented user-centered design is not new method for

application development, it has not previously been applied to natural language processing tasks.

Working translators use a variety of resources that lend themselves to electronic storage and retrieval. These include monolingual and bilingual dictionaries and glossaries, large collections of source and target language text, and other lexical information. Language instructors search through large amounts of text to find authentic examples of language use in particular contexts. Both user groups can benefit from the language analysis tools being developed under the TIPSTER program.

Storage and retrieval of information in electronic databases can be more efficient than equivalent searches in their paper-based counterparts. To be useful for working translators, methods for searching, retrieving, and presenting information must be done in ways that are familiar. Further, the current state-of-the-art in traditional machine translation and programmed instruction provide inadequate support for language translators, learners, and instructors. Although TIPSTER technology does not immediately address these issues, it does provide the basis for new systems that can support humans working with language, and resources that can aid them in their work.

3.0 Development Methodology

There are many problems associated with developing new technology to help with tasks previously solved with old technology. Often, new technology is not delivered in a conveniently usable manner, and systems may not provide functions that are immediately useful to professionals. Usability or usefulness may not be the primary concern of developers of new technology whose attention and creative energies are rightly focused on the mechanisms of the software. Research in human-computer interaction suggests that a user-centered task-oriented approach is the most appropriate method for developing interfaces that deliver new technology to an existing workforce.

CRL has employed a task--oriented user-centered approach to apply natural language technology to the design of interface software that supports working translators, language learners, and instructors. The result of this work are the collection of tools OLEADA. Because the goal has been to get technology into the hands of users in ways that meet their needs, CRL has focused on user testing that motivates feature development and system enhancements.

3.1 Iterative Design and Participatory Prototyping

The design methodology used for developing OLEADA, and its precursor Cibola, is one of iterative design, and the first step in this process is to understand the user through user-protocol task-analysis. User-protocol task-analysis involves an empirical analysis of workers at their jobs, and has three goals. The first goal is to determine worker goals and their strategies for accomplishing them. The second is to characterize workers tasks. The third is to identify cognitive bottlenecks.

Participatory prototyping again focuses on users and their tasks. Here, designers observe users working, and involve users in the stem design process by having them work with early prototypes. User/designer discussions focus on real problems users are having working with the prototypes to accomplish real goals. A process of iterative refinement (user observation, participative-prototyping, and formative evaluations) ultimately shapes useful systems.

Formative evaluations are short, empirical design evaluation studies that focus on system improvement, not system validation. Here, users and developers can identify system problems and enhancements that, if implemented, can significantly improve system usability. Fixing 'details' often leads to expected productivity gains.

3.2 Existing Tools

Many software tools already exist that could be useful to language translators, instructors, and learners. But because they are cumbersome and awkward to learn and use they often go unused by all except computer programmers and developers. Examples include UNIX programs like grep, editors like emacs that can support multilingual text, and programming languages like perl and, lisp, and prolog that can be used to manipulate text data. Most advanced language analysis tools, e.g. name recognizers/taggers, part-of-speech taggers, etc., also demand similar skills. To bring the power of these tools and others to language translators, instructors, and learners requires usable user interfaces that help users accomplish their tasks.

4.0 Four Phases

This section describes three iterations of the four step iterative design process used to develop OLEADA. These steps consist of user observations, task analysis, interface design, and participative prototyping that includes formative evaluations. Typically, research objectives drive research in natural language processing

(NLP). In the context of CRL's work in NLP, however, user observations and task analysis combine to define directions for additional NLP and user interface research and objectives. This is particularly true in the case of OLEADA.

4.1 Phase One

Phase one began with a study (user observation) of DoD analysts engaged in paper and pencil translation tasks. It was observed that translators working with pencil and paper tend to work with a source text alongside the translation in progress rather than above or below it.

During task analysis, it was noted that translators often use a variety of paper-based tools and resources like mono- and bilingual dictionaries, specialized glossaries, and thesauri to aid them in the translation task. These translators would also spend significant amounts of time marking hardcopy, comparing source and target language texts, and consulting reference material.

The initial interface design focused on methods for displaying, editing, and marking-up multilingual text, and the identification of tools and methods for accessing electronic equivalents of paper-based resources. These resources included dictionaries, glossaries, thesauri, and other data including translation memory (parallel-aligned source and target language text).

The first prototype focused on Spanish and English text. The system's user interface technology included a 'bookmark' tool so that users could keep track of searches in reference material, and an annotation tool that enabled users to highlight and attach comments to text. The need to have TIPSTER-style document management for annotations and attributes became apparent early in the user/task analysis. CRL's multi-attributed text widget was used to provide users with facilities for editing and displaying text. NLP technology included a CRL developed Spanish morphology tool to enable sophisticated searches of on-line resources, and used lexical indexing for electronic dictionaries and thesauri. The system also contained a prototype translation memory tool.

4.2 Second Phase

User observation for phase two consisted of users working with the prototype system on example tasks. Task analysis demonstrated that the system lacked functionality and resources. The Spanish morphology component was enhanced, and the alignment algorithm for translation memory was improved. Further improvements were made to the components for multilingual display and edit.

The second prototype contained more resources, had improved document management, and was changed to a client/server architecture. More mono- and bilingual dictionaries were added, as was the CIA Chiefs-of-State database and a world gazetteer. Fuzzy matching was added to aid users in searches.

4.3 Third Phase

The users observed during the third phase were instructors who used the system to develop instructional material. Task analysis of their work showed that a large part of their effort consisted of gathering, retrieving, and analyzing text in context, and current, authentic texts were extremely useful.

Changes to the system included the addition of CRL's concordance tool (X-Concord), improved multilingual information retrieval, and enhanced Chinese/Japanese word identification and segmentation.

5.0 OLEADA

Currently, OLEADA provides a multi-windowed side-by-side presentation of source and target language texts (as opposed, for example, to the top-bottom presentation of texts on some PC's) with editing capabilities in the target text window. To support editing of text displayed in multiple languages, OLEADA utilizes CRL's multi-attributed text widget. This makes it possible for OLEADA users to display and edit multilingual texts. Languages include those using Latin characters, Chinese, Japanese, Arabic, and Russian. OLEADA also uses TUIT, CRL's TIPSTER user interface toolkit, and TDM, CRL's TIPSTER document manager.

5.1 Text Display, Edit, and Annotations

The OLEADA text editor makes it possible for users to display, edit, and annotate multilingual text. During the design process it was discovered that translators often like to make notes on documents they are working on, either in the margins, on the lines themselves, or by attaching notes to the document. To accommodate this a TIPSTER annotation and document management feature was added to the system. Annotated text in Figure 1 appears as colored highlights on the computer screen, or as grayed text here. This feature enables users to make notes or annotations to a document on-line. Annotated text is color-highlighted, has attributes associated with it (such as 'author' or 'type', and can be categorized into groups. Annotations on doc-

uments automatically added by TIPSTER language analysis modules can also be viewed and changed through the same annotation tool.

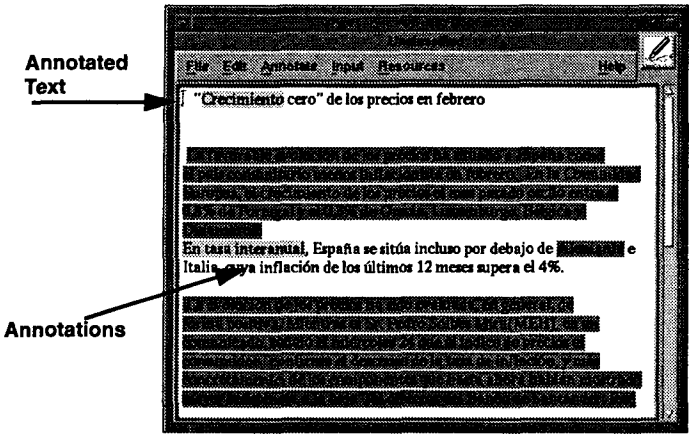

FIGURE 1. Annotated Text

Highlighted text in an OLEADA document is similar to highlighting text with a marker on paper. It draws attention to particular passages. Like writing in the margin of a paper document, OLEADA has an annotation list interface that can be used to associate other text, either as input by the user, or by linking the annotation to text in another document. This feature is shown in Figure 2.

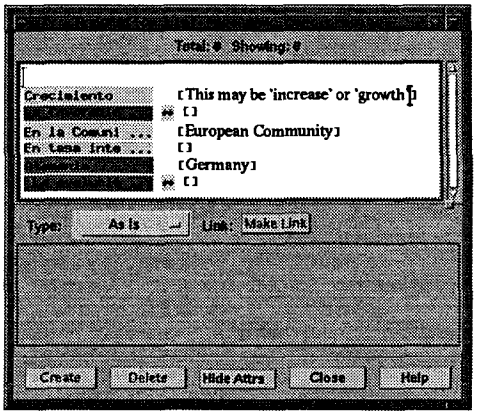

FIGURE 2. The Annotation List Window with Annotation Text

5.2 Xconcord

The XConcord program is a concordance tool that allows KWIC (Key Word In Context) searches to be done in text in as many as 17 languages. It is designed to be easy to work with so that teachers and students can use XConcord in the classroom to identify relevant texts

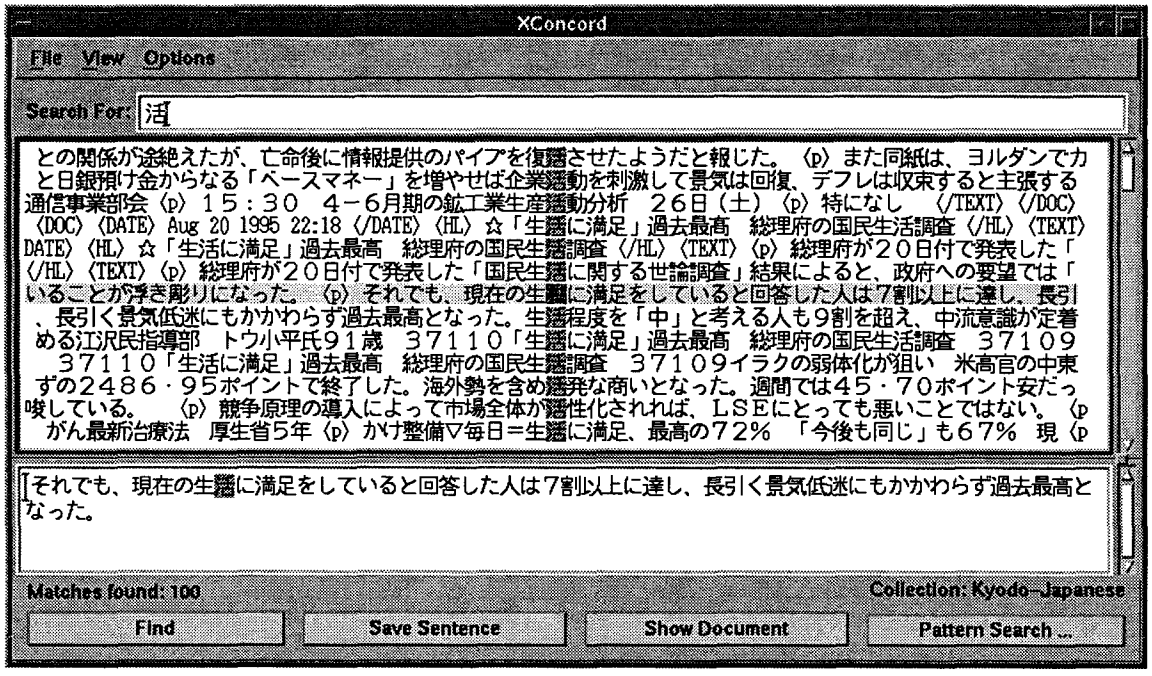

FIGURE 3. The Annotation List Window with Annotation Text

by viewing words and expressions in context. Searching is quick and the size of the corpus is limited only by available disk space. Using an implementation of the Boyer-Moore search algorithm specially adapted for wide characters, X-Concord can search at over 1MB per second, eliminating the need for pre-indexing on many moderate scale corpora.

Searching is very flexible. Users can match any string with any part of a word or phrase. Users can also limit the search to only those concordances either containing or missing specified strings in the context to the left or right of the keyword.

Xconcord shows the results in a KWIC display and also, as seen in the smaller bottom window in Figure 3, the complete sentence for the selected KWIC line. The complete document is displayed in yet another window. Easy methods for saving individual sentences or complete documents to new text files are provided. The users can then edit these files or use Xconcord to print the results.

5.3 The Dictionary Interface

The interface to the OLEADA Dictionary resource includes multilingual access to headwords and their definitions, and also provides users with examples of usage, part-of-speech information, etc. Further, translators often find valuable information by looking at the entries that are close to the target entry. Figure 4 shows the OLEADA Dictionary interface.

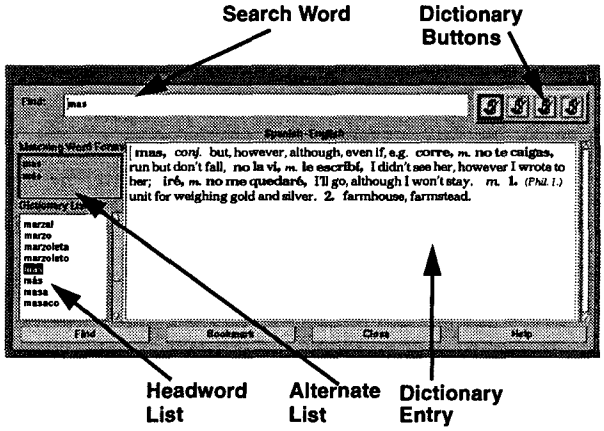

FIGURE 4. The Dictionary Interface

CRL's Dictionary tool provides users with an integrated and easily accessed interface to a wide variety of on-line fixed reference material. Our multilingual, multi-attribute X-Window text capability is used to format this material to capture and reflect the original printed form, complete with all of the lexicographic markup, which makes these on-line resources at least as useful as their printed counterparts. The result is that language translators and learners can use their existing knowledge of how to use these dictionaries. Further, text is fully integrated with the windowing software so it is easy to copy and paste words and phases found in these resources directly into the user's target documents.

A limited view of the headword list is provided. Headword lists consist primarily of the root form of a word, however searches can be performed using morphological variants. OLEADA has morphological analysis component that enables the system to return information on morphological variants of a search term. In cases where similar spelling or morphological variants are available, a fuzzy match list is provided that users can select from.

The Dictionary tool's usability is also enhanced by its sophisticated lexical search capability. Word stemming, Spanish morphology, Chinese and Japanese word segmentation and multiple codeset indexing all help to ensure that every lexical form related to the search term is found. The searches are fast, and all dictionaries are searched each time. This enables users to see which resources have relevant entries. Searches are automatically expanded by a fuzzy matching scheme if the initial search fails. Fuzzy expansion can be helpful in cases where the exact form of a search term is not known, or where you may not recall the spelling of term. A wildcard matching capability is also available.

Dictionary resources in OLEADA are indexed alphabetically by word like their paper counterparts. There are several on-line dictionaries available. During a search all dictionaries are searched. If a match cannot be made in the current dictionary, but occurs in another dictionary that entry will be displayed. The dictionaries are easily-accessible and entries can be retrieved by simply clicking on the dictionary button corresponding to the desired dictionary.

OLEADA's Dictionary interface also provides an alternate word list in addition to the headword list for words with multiple entries that match the morphological form of the search word, such as accented words or words with alternate spellings. Dictionary entries can be retrieved for one of these words by clicking on the word in the list.

Unlike paper dictionaries, OLEADA's on-line dictionaries can be searched using partial words and/or wildcards. A wildcard character is one that can signify any letter and is represented by an asterisk (*). For

example, entries for words beginning with ma but ending with any letter could be found by entering `ma*`. The information returned consists of a fuzzy match list containing all of the words beginning with ma.

5.4 Bookmarks

When using paper-based resources, translators will often make references back to information they have previously found. To accommodate this OLEADA has a 'Bookmark' feature that keeps a list of lexical items previously found. Figure 5 shows the interface for this feature.

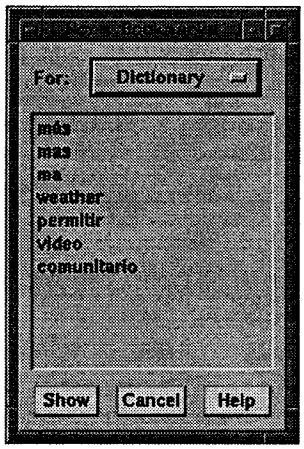

FIGURE 5. The OLEADA Bookmark Window

By selecting a resource and clicking on any of the entries in the search list a user can quickly return to previously retrieved information.

5.5 Word Frequency Tool

CRL's Word Frequency tool provides users with a simple interface for viewing word statistics for individual documents or large collections of documents. In addition, word frequencies in individual documents or smaller sub-collections can be automatically compared to larger collections to identify 'distinctive' words in the document that are significant with respect to the larger collection. This feature can be used to identify important "domain specific" words. By looking at the these frequency lists, a language analyst or instructor can improve their coverage and avoid missing prominent words.

The word frequency tool also works with TIPSTER documents and collections and takes advantage of word segmentation annotations to count Chinese and Japanese words. Documents and collections are processed quickly and results can be re-accessed though collection attributes.

6.0 Conclusion

The benefits of OLEADA are numerous. TIPSTER technology is being transferred in ways that give real users access to technology in useful ways. Machines are excellent tools for quickly searching for, retrieving, and storing information. Humans are good at using language. Through a process of task-oriented user-centered design and iterative refinement, computer tools have been developed that take advantage of the strengths of machines to support the strengths of humans. OLEADA provides users with a consistent, networked medium for working with multilingual text and integrates analysis tools using the TIPSTER architecture.

More importantly, OLEADA offers an informational technology alternative to traditional language instruction. It enables adult professionals, all of whom use informational technology on the job, to access pertinent and authentic materials, perform motivated tasks, and select a range of performance support tools. Learners proceed like researchers as they direct and manage their own training, not only in the classroom but also on the job. Instructors can use OLEADA to support all phases of language training.

7.0 Acknowledgment

CRL's work on OLEADA has been funded by DoD contract #MDA 904-94-C-E086.