# An Integrated System for Source Language Checking, Analysis and Term Management

**Eric Nyberg**
ehn+@cs.cmu.edu

**Teruko Mitamura**
teruko+@cs.cmu.edu

**David Svoboda**
svoboda+@cs.cmu.edu

**Jeongwoo Ko**
jko+@cs.cmu.edu

**Kathryn Baker**
klb+@cs.cmu.edu

**Jeffrey Micher**
jeffreyc+@cs.cmu.edu

Language Technologies Institute
Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA 15213 USA

## Abstract

This paper presents an overview of the tools provided by KANTOO MT system for controlled source language checking, source text analysis, and terminology management. The steps in each process are described, and screen images are provided to illustrate the system architecture and example tool interfaces.

## 1 Introduction

The KANTOO machine translation system combines controlled language checking of source texts with interlingual translation to multiple languages (Nyberg and Mitamura, 2000). The overall KANTOO architecture is shown in Figure 1. This system presentation focuses on the components of the system which are responsible for source language checking, source text analysis, and terminology management. These include the Controlled Language Checker, Analyzer, and Lexical Maintenance Tool. The cyclic workflow for source document creation and translation follows these main steps:

- Document creation, update or reuse. XML documents are created, or existing documents are updated using an XML editor. Portions of existing documents are also copied and pasted to new documents.

- Controlled language checking and issue resolution. The contents of the document are checked for conformance to the controlled language, and any non-conforming sentences are flagged for the author.

- Pending term resolution. Words or phrases which were tagged as new terms during authoring are resolved (accepted or rejected) by the lexicographer; the system's dictionary is updated.

- Document approval. Once all the sentences in the document conform to the controlled language, the document is approved.

- Document translation. Approved documents may be queued for translation to multiple languages.

## 2 Source Language Checking

Figure 2 shows the two views present in the Controlled Language Checker (CLC). The tree view to the left is used to display the internal structure of the XML document. The text view to the right is used to display the content of the currently-selected XML element. During source language checking, the XML tree is traversed, and each checkable segment is broken into appropriate units (sentences, headings, etc.). Each unit is passed to the KANTOO Analyzer module for analysis. If the Analyzer returns a diagnostic message or warning, the corresponding element in the tree view and text view are highlighted appropriately. Any problems which are found must be resolved by the author, either by interacting with the system or by rewriting the text in question.

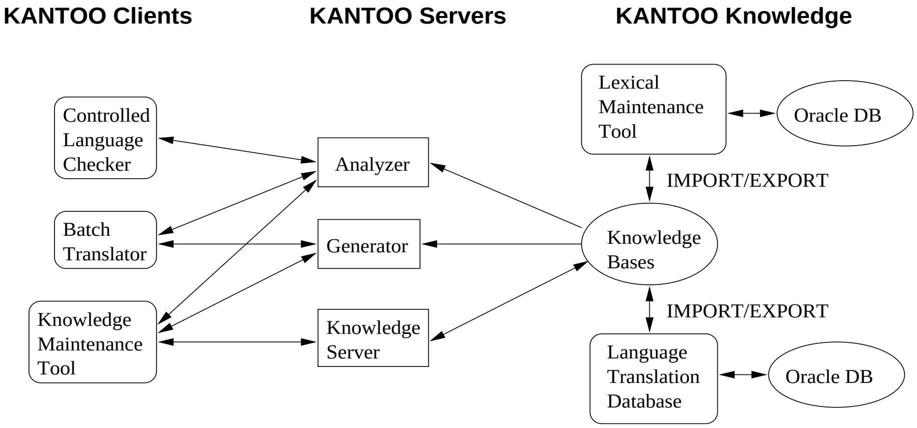
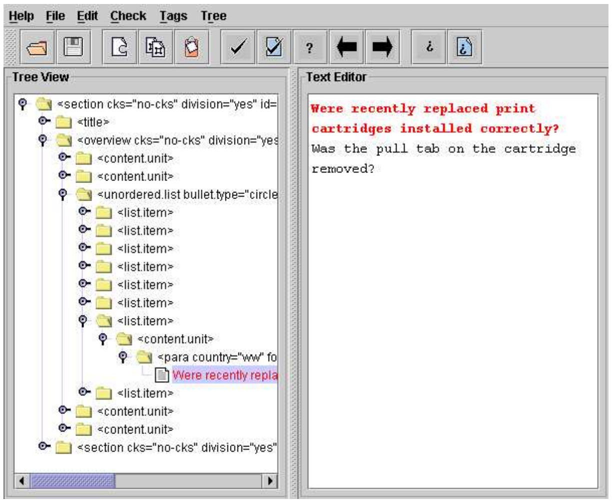


Figure 1: **KANTOO Architecture.**

The system incorporates a variety of diagnostic messages. Some diagnostics include suggested rewrites, which the author can approve with a single mouseclick or keystroke (see Figure 3). Once all of the issues have been resolved, the document is considered "approved" and may be submitted for translation.



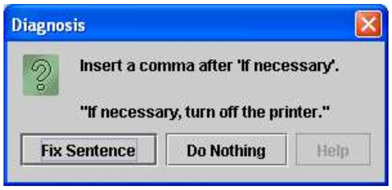
```
<section cks="no-cks" division="yes" id=
  <title>
  <overview cks="no-cks" division="yes
    <content.unit>
    <content.unit>
    <unordered.list bullet.type="circle
      <list.item>
      <list.item>
      <list.item>
      <list.item>
      <list.item>
      <list.item>
      <list.item>
        <content.unit>
          <para country="ww" fo
            Were recently repla
      <list.item>
    <content.unit>
    <content.unit>
<section cks="no-cks" division="yes"
```

```
Were recently replaced print
cartridges installed correctly?
Was the pull tab on the cartridge
removed?
```

Figure 2: **Controlled Language Checker (CLC).**

## 3 Source Language Analysis

The KANTOO Analyzer performs two tasks in the KANTOO system. During document checking, the Analyzer returns specific diagnostic messages for sentences which require system and/or author intervention. During document translation, the Analyzer produces an interlingua representation for each input unit in each of the checkable segments of the document. The Analyzer performs the following steps for each input:

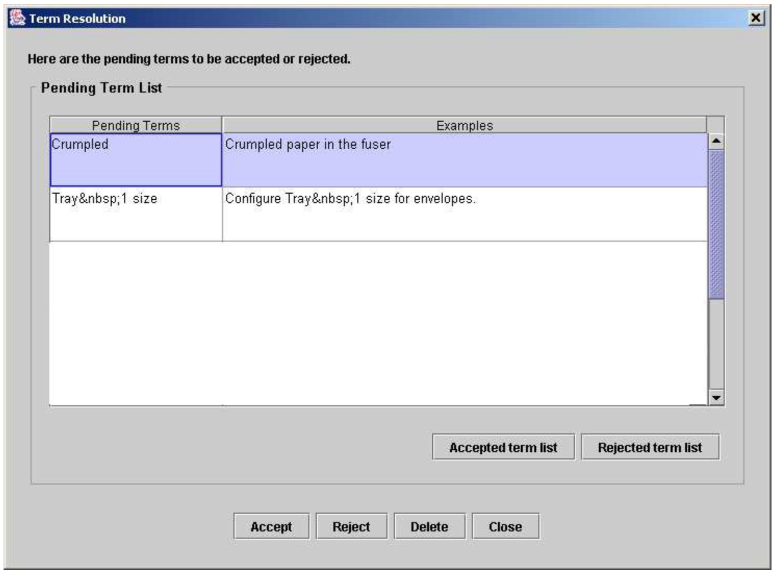


Insert a comma after 'If necessary'.

"If necessary, turn off the printer."

Figure 3: **Interactive Rewriting.**

- Segmentation. Translation units (sentences, headings, etc.) are identified. Within each unit, individual tags and tokens are identified.
- Lexicalization. Input tokens are associated with known entries in the system's dictionary.
- Syntactic Parsing w/ Diagnostics. Each translation unit is parsed using a syntactic grammar to check for controlled language conformance. Specific diagnostic messages may be generated for each sentence.
- Automatic and Interactive Disambiguation. Translation units with more than one legal parse are resolved, either via a) automatic selection of the most likely meaning, using a set of disambiguation heuristics, or b) interactive clarification with the user.
- Semantic Interpretation. An interlingua expression is associated with each translation unit.



Term Resolution

Here are the pending terms to be accepted or rejected.

Pending Term List

| Pending Terms | Examples |
|---|---|
| Crumpled | Crumpled paper in the fuser |
| Tray 1 size | Configure Tray 1 size for envelopes. |

Figure 4: **LMT Pending Terms Table.**

- Pattern-based Diagnostics (for failed parses). When a sentence fails to parse, a set of pattern-based diagnostics is used to search the original token sequence for possible problems.

## 4 Terminology Management

Terminology management is a cyclic process that integrates document authoring and dictionary update:

- New terms are marked with `<AddTerm>`. A special XML tag is used to mark terms that aren't in the dictionary, but which the author feels should be added. The CLC allows these terms to pass.

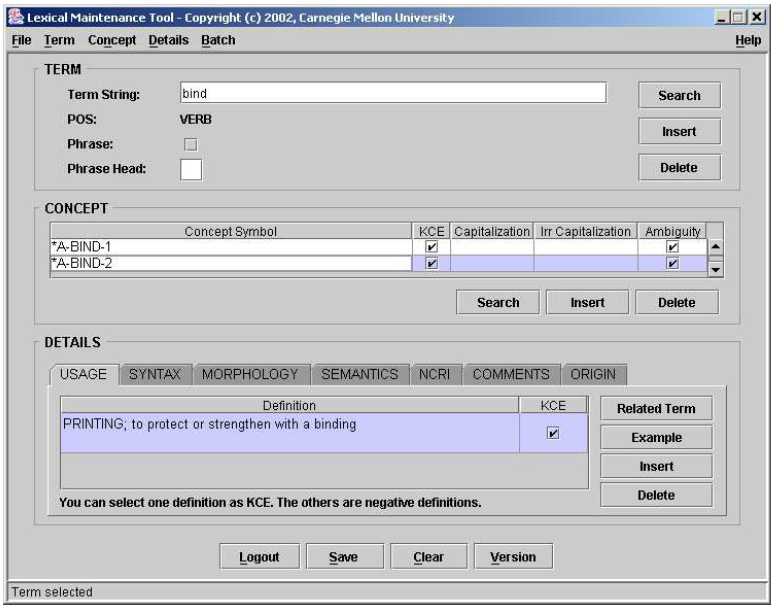
- Pending terms are added to the LMT database. When the CLC encounters a term that is not in the dictionary and is tagged with `<AddTerm>`, it adds a new row in the pending terms table maintained by the Lexical Maintenance Tool (LMT). See Figure 4.

- Lexicographer resolves pending terms (accept, reject). Using the LMT, the lexicographer either accepts or rejects each pending term. For newly accepted terms, the system will create a default entry based on the part of speech; the lexicographer can use the main LMT window ( see Figure 5) to edit the default entry and add other information.

- `<AddTerm>` tags are re-checked. Once the new terms have been accepted or rejected, they can no longer appear inside `<AddTerm>` tags. When it encounters a term inside an `<AddTerm>` tag, the CLC check to see whether the term has already been accepted or rejected. In either case, it disallows the use of the `<AddTerm>` tag.

## 5 Deployment

The KANTOO CLC is currently implemented as a small Java client program that runs on the author's desktop. The KANTOO Analyzer is implemented in C++, and runs as a network server shared by several authors simultaneously. The LMT database also runs as a network server, and can support connections from multiple authoring sessions simultaneously. The LMT interface runs as a Java applet or desktop application, and can be remotely accessed by the lexicographer. The KANTOO CLC also includes comprehensive, self-paced training and assessment materials which are accessed on-line via a web browser. The KANTOO tools also incorporate extensive on-line help.



Term String: bind

POS: VERB

Phrase: □

Phrase Head:

| Concept Symbol | KCE | Capitalization | Irr Capitalization | Ambiguity |
|---|---|---|---|---|
| *A-BIND-1 | ☑ | | | ☑ |
| *A-BIND-2 | ☑ | | | ☑ |

| Definition | KCE |
|---|---|
| PRINTING; to protect or strengthen with a binding | ☑ |

You can select one definition as KCE. The others are negative definitions.

Term selected

Figure 5: **LMT Main Window.**

## Bibliography

Kamprath, C., E. Adolphson, T. Mitamura and E. Nyberg (1998). “Controlled Language for Multilingual Document Production: Experience with Caterpillar Technical English,” *Proceedings of CLAW 1998*, Pittsburgh.

Mitamura, T., Nyberg, E. and Carbonell, J. (1991). “An Efficient Interlingua Translation System for Multi-lingual Document Production,” *Proceedings of Machine Translation Summit III*, Washington, D.C.

Mitamura T., E. Nyberg, E. Torrejon, D. Svoboda, A. Brunner and K. Baker (2002). “Pronominal Anaphora Resolution in the KANTOO Multilingual Machine Translation System,” *Proceedings of TMI-2002*, March 2002, Keihanna, Japan.

Nyberg, E. and T. Mitamura (2000). “The KANTOO Machine Translation Environment,” Proceedings of AMTA-2000, Cuernavaca, Mexico.

Nyberg, E., T. Mitamura and W. Huijsen (2003). “Controlled Language,” in H. Somers, ed., *Computers and Translation: Handbook for Translators*, Johns Benjamins.