# Automated Corpus Analysis and the Acquisition of Large, Multi-Lingual Knowledge Bases for MT[1]

Teruko Mitamura
Eric H. Nyberg, 3rd
Jaime G. Carbonell
Center for Machine Translation
Carnegie Mellon University
Pittsburgh, PA 15213

**Abstract**

Although knowledge-based MT systems have the potential to achieve high translation accuracy, each successful application system requires a large amount of hand-coded knowledge (lexicons, grammars, mapping rules, etc.). Systems like KBMT-89 and its descendents have demonstrated how knowledge-based translation can produce good results in technical domains with tractable domain semantics. Nevertheless, the cost of developing large-scale applications with tens of thousands of domain concepts precludes a purely hand-crafted approach. The current challenge for the "next generation" of knowledge-based MT systems is to utilize on-line textual resources and corpus analysis software in order to automate the most laborious aspects of the knowledge acquisition process. This partial automation can in turn maximize the productivity of human knowledge engineers and help to make large-scale applications of knowledge-based MT an economic reality. In this paper we discuss the corpus-based knowledge acquisition methodology used in KANT, a knowledge-based translation system for multi-lingual document production. This methodology can be generalized beyond the KANT interlingua approach for use with any system that requires similar kinds of knowledge.

## 1. Introduction

Recent work in the field of Knowledge-Based Machine Translation has demonstrated that the use of explicit representations of domain words and concepts along with semantic processing during translation can achieve a high degree of accuracy in the target text (Goodman and Nirenburg, 1991). In particular, prototype development within the KANT architecture (Mitamura, Nyberg & Carbonell, 1991; Nyberg & Mitamura, 1992; Carbonell, Mitamura and Nyberg, 1992) has shown that current knowledge-based techniques can achieve high accuracy and high translation throughput rates for multi-lingual translation, when applied to a controlled source vocabulary and grammar in a restricted technical domain.

In the evolution from prototype to full system, we have focused on scaling up the KANT system in a large, practical application for multi-lingual translation of heavy equipment manuals[2]. Many MT researchers feel that the primary challenge faced by the knowledge-based approach is the difficulty of building enough knowledge for a system of practical size, especially if the creation of that knowledge is complex and labor-intensive. We have accepted this challenge, and we acknowledge

---

[1] Proceedings of the 5th International Conference on Theoretical and Methodological Issues in Machine Translation, Kyoto, Japan, July 14-16, 1993.

[2] The current KANT application domain is the full line of Caterpillar products, and the task is translation from controlled English source to multiple target languages without post editing. The project is sponsored by Caterpillar and implemented by the Center for Machine Translation at CMU in conjunction with Carnegie Group, Inc.

that large knowledge bases cannot be constructed entirely by hand. As stated in (Carbonell, Mitamura & Nyberg, 1992), we firmly believe that the analysis of corpora as a knowledge resource is indispensable in building knowledge-based MT systems. The focus is not on the use of corpus examples at run time, as in example-based translation (Sato, 1991), but on the extensive use of corpus examples as an off-line resource used in the automated creation of run-time knowledge sources.

This type of knowledge acquisition methodology is part of the next generation in knowledge-based MT systems – application systems for large technical domains which leverage existing text resources to make system development economically feasible. The idea of deriving domain knowledge from corpora is not a new one; see, for example, (Grishman & Kittredge, 1986; Grishman & Kosaka, 1992; Zernik, 1991). Nevertheless, the current KANT system represents one of the first attempts to incorporate this kind of next-generation knowledge acquisition methodology into a large-scale practical MT development project.

In this paper, we begin with a brief summary of the knowledge sources used in the KANT system. In the sections that follow, we describe the acquisition of knowledge sources using on-line texts as a primary resource; for each knowledge source, we describe our methodology and results, and then discuss the implications for both knowledge-based MT system development and MT systems in general. Our focus is not on tools for end-user customization of the system (for example, entering a custom dictionary). In contrast, knowledge acquisition tools are used by the systems engineer in order to support creation of not only lexicons, but also grammars, mapping rules, domain concepts, etc. The techniques used in KANT can be utilized in any domain or for any source and target languages. Although specifically applied to interlingual knowledge-based translation, these techniques can be generalized to any MT system which requires similar types of domain knowledge, such as semantic-transfer approaches.

## 2. Overview of the KANT System

The basic architecture of KANT is shown in Figure 1. The system makes use of the following knowledge sources:

- A Source Grammar for the input language which builds syntactic constructions from input sentences;
- A Source Lexicon which captures all of the allowable vocabulary in the domain;
- Source Mapping Rules which indicate how syntactic heads and grammatical functions in the source language are mapped onto domain concepts and semantic roles in the interlingua;
- A Domain Model which defines the classes of domain concepts and restricts the fillers of semantic roles for each class;
- Target Mapping Rules which indicate how domain concepts and semantic roles in the interlingua are mapped onto syntactic heads and grammatical functions in the target language;
- A Target Lexicon which contains appropriate target lexemes for each domain concept;
- A Target Grammar for the target language which realizes target syntactic constructions as linearized output sentences.

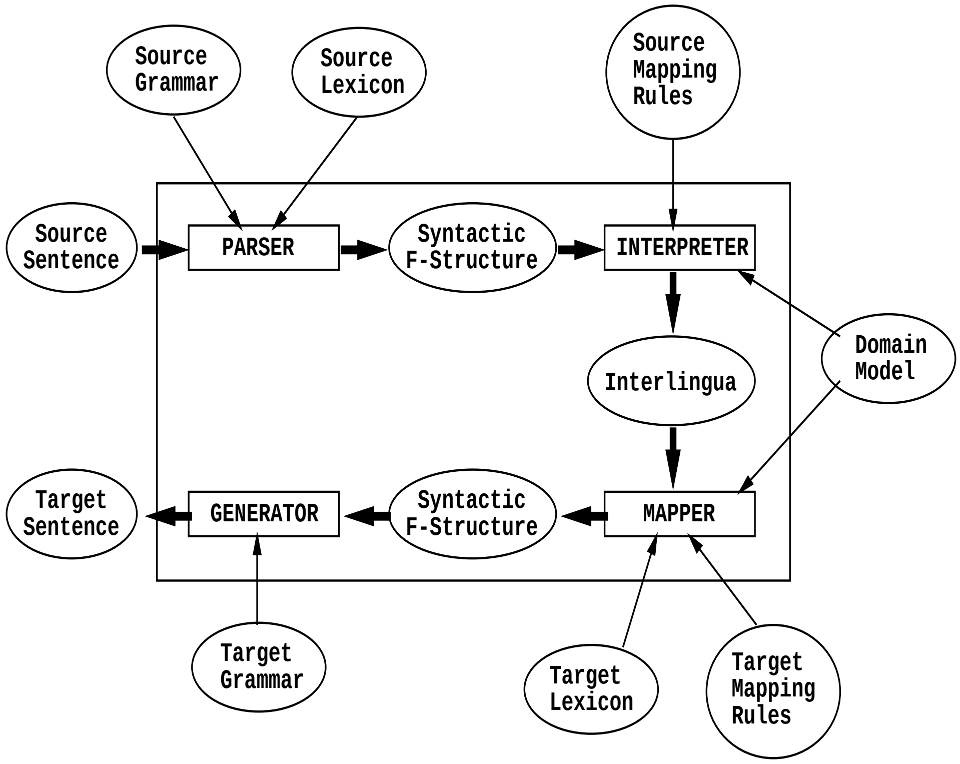


Figure 1: **The Run-Time Architecture of KANT**

Because the focus of this paper is on knowledge acquisition rather than the internal details of the run-time system, the reader is spared further detail concerning the Parser, Interpreter, Mapper and Generator. More detail can be found in (Mitamura, Nyberg and Carbonell, 1991) and (Nyberg and Mitamura, 1992).

The relationships between text resources (corpora) and the development of KANT knowledge sources is shown in Figure 2. Starting with the original source and target language corpora, the goal of knowledge acquisition is to continually re-use existing resources in creating additional resources and knowledge in a "value-added" way. For example, the source corpus is used to derive a source lexicon and mapping rules, which are in turn used to derive the knowledge in the domain model. In the remainder of this paper, we discuss the individual steps in creating KANT knowledge sources from corpora.

## 3. Source Language Lexicon

In creating a KANT application for a particular domain, the first task is to create a source vocabulary for domain text. This is achieved by first identifying a comprehensive set of documents covering the whole domain (the "Raw Corpus") which is to be analyzed for extraction of vocabulary items. Then a set of automated procedures are used to extract candidate vocabulary entries, which is then subjected to human refinement.

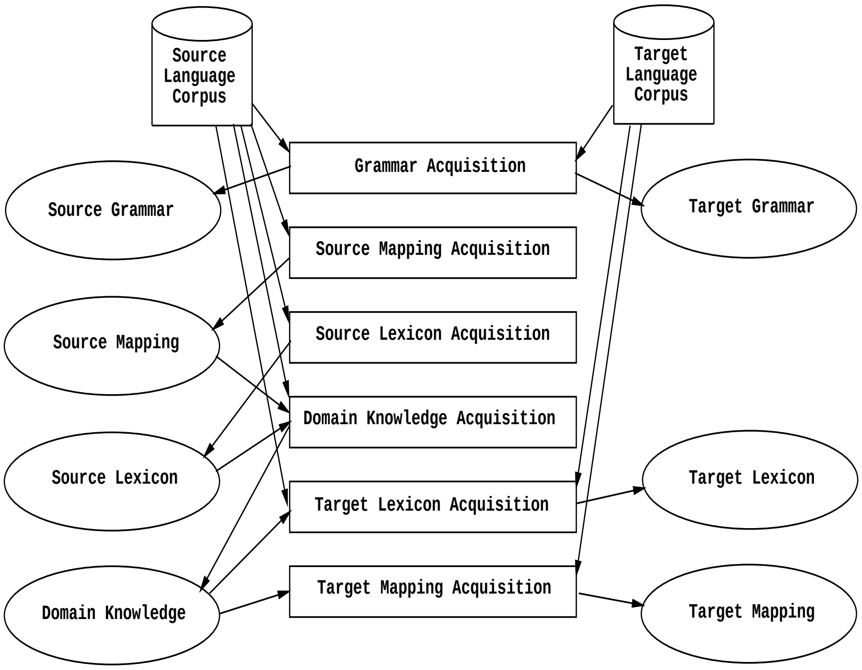


Figure 2: **Knowledge Acquisition in KANT**

### 3.1. Methodology

The steps taken to construct a lexicon from the source corpus are as follows (cf. Figure 3):

1. *Automatic Deformatting of the Raw Corpus*. The raw corpus is processed by a set of programs which remove and/or canonicalize the formatting codes used in the source documents.

2. *Automatic Creation of a Word Corpus*. All occurrences of inflected forms are counted and merged into a corpus of word occurrences by a statistical program.

3. *Automatic Creation of a Sentence Corpus*. All of the sentences which appear in the corpus are indexed by the words that appear in them, in order to support further analysis, including KWIC (Key Word in Context) access to the corpus.

4. *Creation of the Initial Word and Phrase Lexicons*.

   In order to produce an Initial Lexicon, the Lexicon Creation Program uses a Tagged Corpus (e.g., the tagged Brown Corpus (Francis & Kuĉera, 1982)) as a resource for part-of-speech information, in conjunction with a source language morphological analyzer. The Initial Lexicon contains a part of speech marker for each root form found in the Word Corpus. In order to produce the Initial Phrasal Lexicon, the Phrasal Lexicon Creation Program uses both the Initial Lexicon and the Sentence Corpus as resources (see Figure 4 for examples of words and phrases).

5. *Human Refinement of the Lexicons*.

   The Initial Lexicon and Initial Phrasal Lexicon are refined by a process of human inspection, which includes use of the Sentence Corpus via a KWIC browsing interface.

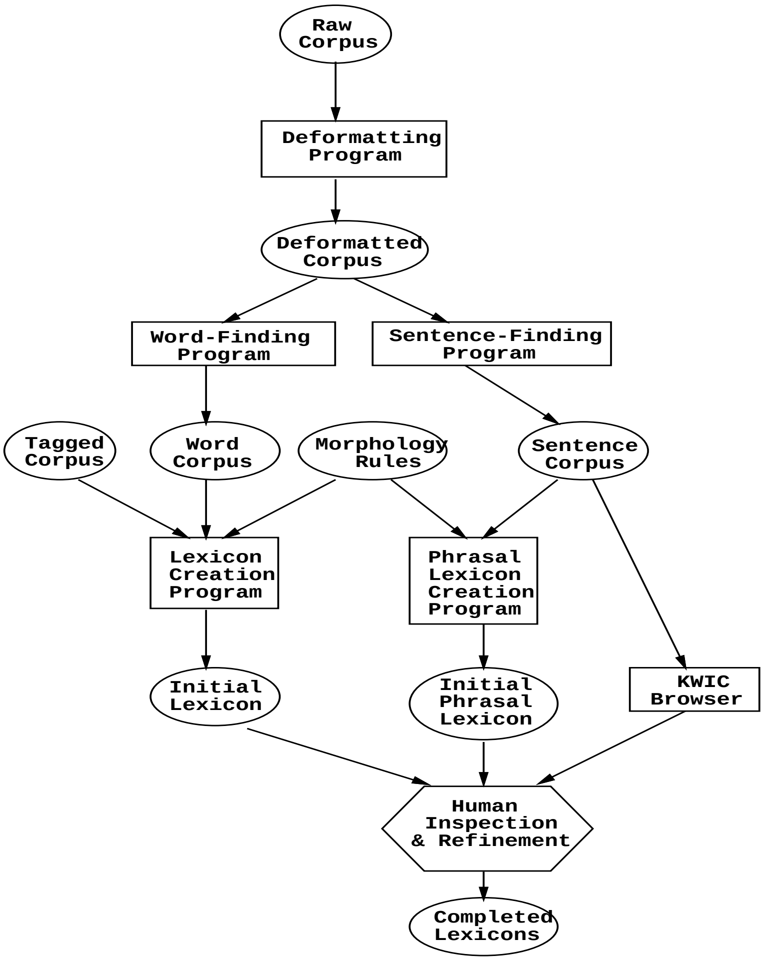


Figure 3: **Automation of Source Lexicon Acquisition**

| | | | |
|---|---|---|---|
| 630186 | the | 6092 | lb ft |
| 185352 | and | 4889 | cooling system |
| 178946 | to | 3966 | fuel injection |
| 140407 | of | 3862 | parking brake |
| 127776 | in | 3410 | relief valve |
| 103206 | is | 2926 | cylinder head |
| 95967 | a | 2924 | test force |
| 69145 | valve | 2886 | oil pressure |
| 68193 | for | 2789 | control valve |
| 65787 | oil | 2634 | torque converter |
| 63046 | on | 2587 | service hours |
| 56781 | engine | 2491 | o-ring seal |

Figure 4: **Most Frequent Fragments: Words and Phrases in Initial Lexicons**

```
((:ROOT "rip")
 (:POS V)
 (:CONCEPT *A-RIP)
 (:SYL-DOUBLE +)
 (:SYN-FEATURES (VALENCY TRANS INTRANS))
 (:NOTE
  (:SENSE
     "Technical term: to slash into with a ripper"
     "There are several ways to rip hard spots
      and boulders."
     "Rip downhill whenever possible."
     "Do not rip and doze at the same time."))
 (:FREQUENCY 106 368)
 (:UPDATED (20 29 18 26 6 1992) "ehn"))
```

Figure 5: **Example Lexicon Entry**

An example of a finished lexicon entry is shown in Figure 5. The `:ROOT`, `:POS`, `:CONCEPT`, `:SYL-DOUBLE`, `:SYN-FEATURES` and `:FREQ` fields are created automatically with default values. Subsequently, the lexicographer browses occurrences of the word in the Sentence Corpus using a KWIC browser, and refines the default values and also adds a definition and examples to the `:SENSE` field. These are not intended for use by the system, but are provided as a resource for future human readers of the lexicon.

### 3.2. Results

This method has been applied to a large corpus of technical manuals for heavy equipment consisting of approximately 53 megabytes of deformatted text. A single-word lexicon of about 9000 items was extracted from the corpus. A phrasal lexicon of about 50,000 phrases was also extracted from the corpus.

The Tagged Corpus contains some part of speech assignments which are not part of the application domain[3]. Because the phrase-finding heuristics used by the Phrasal Lexicon Creation Program identify noun phrases by searching for strings of adjectives and nouns, some words which are not adjectives or nouns in the domain turn up in phrases in the Initial Phrasal lexicon. Some examples are shown in Figure 6. For example, the word “still” is not a noun in the heavy equipment domain; however, since the Tagged Corpus contains a noun part of speech tag for “still”, phrases like “actuator still” are erroneously placed in the Initial Phrase Lexicon. For this reason, it is important that the Initial Phrase Lexicon is updated once the refinement of the Initial Lexicon has taken place. This is performed by automatically extracting those phrases which contain words whose refined part of speech no longer allows them to participate in phrases. The part-of-speech assignments in the Completed Lexicon are much narrower than those in the Tagged Corpus and cover precisely the usage found in the domain.

[3] A tagged version of the Brown Corpus was utilized as a resource for part-of-speech tags.

```
(JOG ENGINE STARTER)               ;; "jog" not a noun in the domain
(ANNUNCIATE OVERCRANK SHUTDOWN) ;; "annunciate" not a noun in the domain
(LINE BEHIND TRACTOR)              ;; "behind" not a noun in the domain
(ACTUATOR STILL)                   ;; "still" not a noun in the domain
```

Figure 6: **Example Phrase Refinements**

### 3.3. Discussion

It is much more cost-effective to utilize lexicographer time in the refinement of lexicon entries (addition of sense information, examples, etc.) rather than in the creation of the entire lexicon from scratch. Time which would normally be spent in typing in the entries by hand can be spent browsing the corpus of examples in order to refine the meaning of the word to the appropriate domain reading. This technique should be useful for any MT system which is faced with the task of creating lexicons from source corpora.

One problem is that generally-available lexical resources are not always appropriate for narrow technical domains. Our experience with using a tagged corpus is that it is a good resource for “bootstrapping” an initial lexicon, but that there is a significant effort required to narrow the meanings of words which have many general senses but only a narrow technical sense in the domain. Future availability of refined technical lexicons for specific domains should measurably improve the results of this acquisition methodology, while simultaneously reducing the cost of human refinement of lexical entries. We would also like to investigate the potential use of automatic tagging (for example, using an algorithm like Church’s algorithm) as an alternative source of POS assignments.

Another problem that we faced is that the phrase finding heuristics sometimes missed domain noun phrases because they contained words that would normally not be considered part of a noun phrase in English. For example, many `-ing` forms (e.g., as in *summing valve, steering wheel*) denote processes and are nominal in nature, and must be entered in the lexicon as such. Once these classes of missing terms were identified, the phrase-finding heuristics were adjusted and the corpus was re-analyzed to detect them.

Despite the problems we encountered, this methodology proved invaluable in that it supported the creation of an initial lexicon with reduced human effort, thereby supporting the completion of a finished lexicon of 60,000 words and phrases using a practically feasible amount of effort.

## 4. Target Language Lexicon

Once a source vocabulary has been constructed for the domain, the corresponding target language vocabulary must also be constructed for each target language. Because the typical KANT application is for a narrow technical domain (like heavy equipment or software terminology), often with a well-defined customer style in each target language, a domain translation expert is usually consulted. Nevertheless, a significant amount of work is accomplished automatically or through the efforts of personnel who are skilled in the target language but not domain experts.

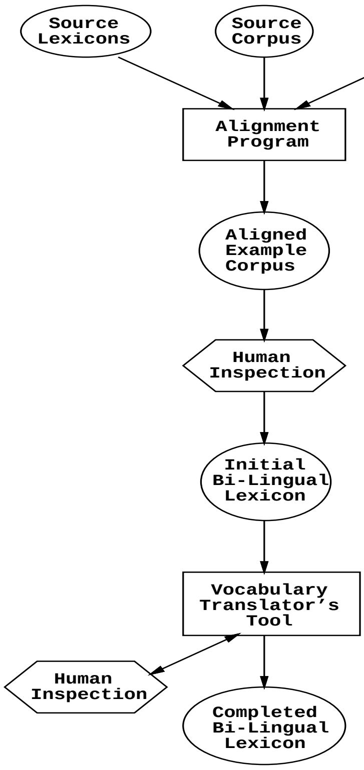


Figure 7: **Automation of Target Lexicon Acquisition**

### 4.1. Methodology

The steps taken to construct a target language lexicon are as follows (cf. Figure 7):

1. *Creation of an Aligned Example Corpus*.

   Sets of corresponding source/target manuals are analyzed in order to locate sections of target text that should contain the translation of particular lexical items in the source language. This is accomplished using the source lexicon to access source words, the source corpus to locate contexts in source texts where those words appear, and an alignment program to locate contexts in target texts which match the source contexts. The alignment is based on cues like document markup and formatting codes, as well as information about correspondences between general words and domain-specific words (like serial numbers) in the source and target languages. Since technical documentation is highly structured (containing many figure references, sectioning commands, etc.), it is often possible to locate the precise target paragraph which corresponds to the source paragraph where a word or phrase appears in the correct context.

2. *Browsing the Aligned Corpus for Translations*.

   A process of human inspection, using a bi-lingual corpus browser, is used to select the actual translation of the source items. A two-window display is used in which the left window contains example paragraphs from the source text, with a particular source lexical item highlighted; the right window contains corresponding example paragraphs from the target text. The human operator selects the appropriate translation from the target text using a mouse, or skips the example if it does not contain an appropriate translation of the source term. For example, Figure 8 shows an aligned context for a paragraph containing the term “ripper tip”; following visual inspection of the French context, the operator can highlight “pointe de ripper” with the mouse and store this as one translation for “ripper tip”. A browsing process like this is carried out for all aligned contexts which exist for any of the source terms which appear within the aligned corpus. Because the corpus will contain many contexts in which different target language terms are used for the same source language term, the tool keeps track of multiple selections in the target context. As a result of this browsing process, the Initial Bi-Lingual Lexicon is created. Further refinement of this process is under investigation.

Source:
“when ordering the new 6y9473 adapter for the former shank assembly , the retainer pin requirements change . the replacement adapter is counterbored . two 9n4245 pins will be needed for the reworked shank . one for the lower hole of the protector and other for the **ripper tip**. two 6y1205 retainer assemblies hold the two 9n4245 pins in position . a 6y2443 pin assembly must be used in the upper protector hole.”

Target:
“une nouvelle clavette est nécessaire lorsque l’on monte le nouveau port-pointe 6y9473 sur l’ancienne dent . le port-pointe comporte des chambrages. deux clavettes 9n4245 seront requises pour la dent modifiée , une pour le trou inférieur du protecteur et l’autre pour la **pointe de ripper** . deux arrêtoirs 6y1205 maintiennent les deux clavettes 9n4245 . il faut utiliser une clavette 6y2443 dans le trou supérieur du protecteur . pour les pièces requises , se reporter au tableau approprié .”

Figure 8: **Example Alignment of Source and Target Contexts**

3. *Vocabulary Translation*.

   At this stage, the expert translator reviews the contents of the initial bi-lingual lexicon, and adds translations for any source terms which have not yet been translated. This is accomplished using a vocabulary translator’s tool which displays terms and their translations (cf. Figure 9). The Tool also provides a streamlined interface which allows the re-use of previous translations. As each term is translated, a “draft” facility is provided which searches for similar terms that have already been translated and places the translation of the nearest match into the translation buffer, thus minimizing the amount of typing necessary. For example, in Figure 9 the family of terms containing “charge valve” is shown at the point where the translator has just finished translating the last term in the family, “nitrogen charge valve”. In order to accomplish this, the translator might have pushed the “Draft” button when “nitrogen charge valve” came up to be translated; in this case the Tool would have placed “soupape de charge” (the translation of “charge valve”, the nearest match) into the translations buffer. Hence the translator need only add any missing material to a previous kernel translation. By ordering the terms to be translated according to increasing length (for example, by creating families of terms which contain a kernel 2-word term, and sorting them in dictionary order), the efficacy of the draft facility is maximized.

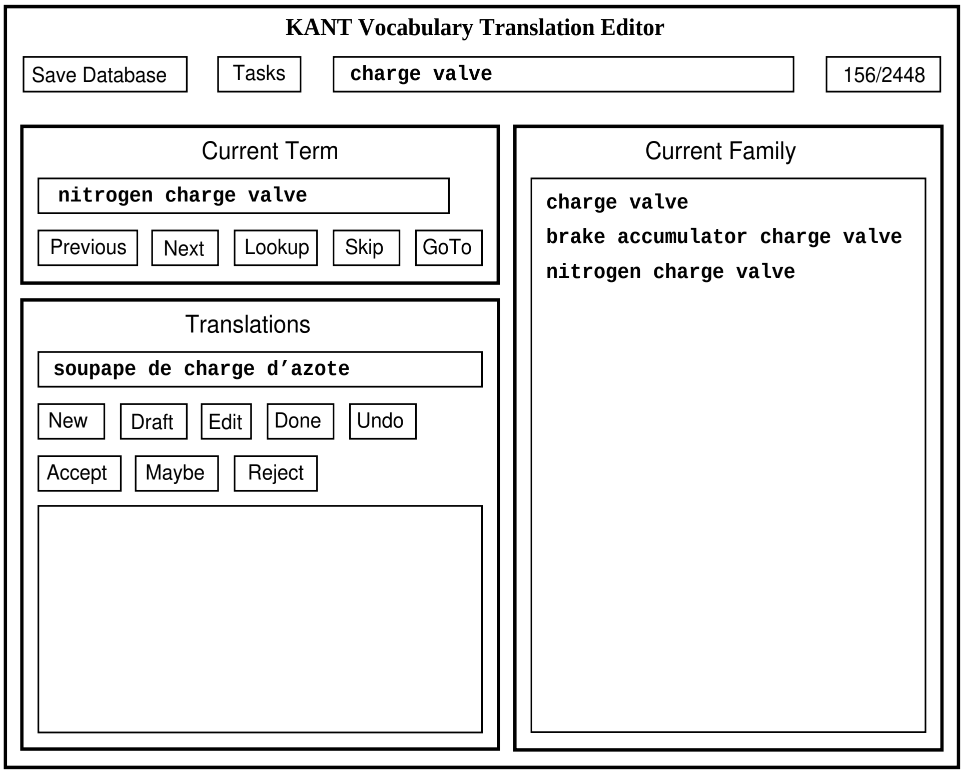


Figure 9: **KANT Vocabulary Translation Editor**

### 4.2. Results

We have successfully employed this methodology in the creation of an English/French bilingual lexicon. Because our domain corpus did not contain a majority of corresponding documents in both English and French, we were only able to extract aligned examples for about 12% of the source lexicon. The initial bi-lingual lexicon was extracted from these examples in just a few weeks by undergraduates with a background in French but with no expertise in the domain. The completion of the bi-lingual lexicon by the domain translation expert is currently in progress. Using the vocabulary translator's tool, along with the pre-existing translations from the initial bi-lingual lexicon and the tool's draft facility, the expert translator is able to produce entries on-line in a format later used to automatically create source and target lexicons. This is much more efficient than manual translation plus transcription or reformatting electronic entries in non-compatible formats.

### 4.3. Discussion

This methodology for target language lexicon creation has three important characteristics:

- Existing domain text provides an important resource for lexicon creation, which reduces human effort by providing target language contexts from which translations can be selected easily by cut-and-paste methods;

- The use of aligned domain text allows some translation work to be done by those skilled in the target language but not skilled in the domain;
- The use of pre-translated terms and a tool which supports transparent re-use of previously translated material maximizes the productivity of the expert translator when working on the remaining terms to be translated.

These techniques provide a resource for system developers and linguists which is useful for any system that involves the creation of a target lexicon or bi-lingual lexicon.

## 5. Source and Target Language Grammar

The input text to KANT is grammatically controlled (Nyberg & Mitamura, 1992). KANT limits the use of constructions that would create unnecessary ambiguity or other difficulties in parsing, while still providing the author with a sublanguage which is sufficient for authoring of clear, understandable technical texts.

Since the source language corpus is not usually controlled, a specification for the controlled grammar (CG) must be defined for the particular application domain. Analysis of the corpus plays a crucial role, since the CG should reflect the style of domain documents and promote author habitability and system acceptance. A separate system is used to interface with the author to ensure compliance with KANT's grammatical restrictions. For the present application, this function is fulfilled by the Language Environment developed by Carnegie Group, Inc. for Caterpillar, Inc.

### 5.1. Methodology

The steps taken to construct a source and target language grammar are as follows:

1. *Examination of the Sentence Corpus*.

   Since the Sentence Corpus contains various types of sentence structures, some of which may be problematic for parsing, we first need to identify the different types of structures in the source language corpus. We list the sentence types which may cause unnecessary syntactic ambiguity and/or increased computational complexity during parsing. Such sentence types include coordination of verb phrases, some types of relative clauses, and various kinds of infinitival clauses.

   We tentatively rule out the complex structures and rewrite these sentences into less problematic constructions. For example, complex coordinate constructions are rewritten into several separate sentences.

2. *Parsing a Sample Sentence Corpus*.

   We construct a sample sentence corpus which includes examples of the different types of constructions in CG. In addition, the lexicographers create usage examples for each of the general words (e.g., verbs) in the source lexicon during lexicon refinement. Currently, there are over 10,000 sentences in the Sample Sentence Corpus. The Sample Sentence Corpus is parsed with an existing source grammar in order to test the complexity of the sentences. A user-oriented definition of the controlled source language (the CG Specification) is written based on the Sample Sentence Corpus.

3. *Refining a Controlled Source Grammar*.

   Based on the CG Specification, we refine the existing source grammar to make sure that all the types of constructions specified in CG will be parsed, while non-CG constructions will be prevented from parsing. Once the CG Specification has been used to train domain authors, new texts can be constructed in the controlled source language which can then be used to further test grammatical coverage, source language style, etc. This operational filter for CG is incorporated into a document authoring tool such as the one developed by Carnegie Group for Caterpillar.

4. *Defining a Target Grammar*.

   The set of syntactic constructions to be used in the target text is based on: a) the CG Specification for the source language, which indicates what kinds of sentences are allowed in the domain, b) the Interlingua Specification, which defines the input to the generation phase and indicates how the semantics of the input will be represented, and c) a set of indexed corpora for the source and target languages. As previously discussed in Sections 3 and 4, source and target lexicon creation produces a KWIC browser for both source and target languages, along with a set of aligned contexts which illustrate how source words and their target translations appear in their respective corpora. This provides a valuable resource for the linguist who must ensure that the target grammar supports all of the syntactic structures associated with the target text.

### 5.2. Results and Discussion

In research-oriented prototypes, the domain is often limited to a small set of input sentences, sometimes just a few hundred sentences. In such cases, a grammar can be written without the use of on-line corpora. In scaling up a system to a large-scale practical domain, it becomes much more important to anticipate and account for all the potential structures that may arise in the domain; in this situation, a large body of historic documents is a valuable resource. The result is that an initial prototype grammar can be extended for much broader coverage once these resources become available.

Because the KWIC browsing utilities and aligned corpora are necessary for lexical acquisition, it is fortuitous that they may also be used for the subsequent process of grammar construction. This helps to maximize the reuse of resources which have a significant amount of added value. Nevertheless, the process of grammar extension and refinement resists automation and remains primarily a task for the human linguistic expert.

## 6. Source and Target Language Mapping Rules

In KANT, source language sentences are syntactically analyzed and semantically interpreted, and the resulting interlingua representation is used as an intermediate stage in multi-lingual translation.

There are two types of source language mapping rules used in KANT: Lexical Mapping rules and Argument Mapping rules (Mitamura, 1989). Lexical Mapping rules map a word and part of speech (e.g., “gas”, N) onto a set of domain concepts (e.g., *O-GASOLINE, *O-NATURAL-GAS). In KANT, lexical mappings are pointers to leaf concept nodes, which are created automatically through application of rules for each type of word (e.g. Noun, Adj).

Because the KANT source language syntactic grammar is written in an LFG-style formalism (Goodman and Nirenburg, 1991), the Parser produces an f-structure representation of the grammatical functions in the sentence (e.g., SUBJECT, OBJECT). The Interpreter then maps each grammatical function onto an appropriate semantic role in the interlingua. Complexity arises when one type of grammatical function can map onto more than one semantic role, depending on the type of semantic head and/or the type of role filler. In these cases, the system must use restrictive mapping rules which license only those syntactic attachments which correspond to the correct assignment of semantic role.

### 6.1. Methodology: Source Language Mapping Rules

The construction of source language mapping rules require resolution of attachment ambiguities and role assignment ambiguities. The steps to construct these mapping rules are as follows:

1. *Identify Set of Domain Semantic Roles*.

   (a) For each type of syntactic attachment which is associated with semantic role assignment (e.g., VP + PP, NP + PP), we build a syntactic pattern and extract all example sentences from the corpus which match the pattern.

   (b) The sentences are grouped according to the attached argument; for example, from the set of sentences which match VP + PP and NP + PP, we would extract sets of sentences which contain each of the prepositions in the domain.

   (c) For each possible meaning of the attached argument, a semantic role is created and a canonical example is produced:

   ```
   LOCATED-BELOW   "Remove the floor plate below the operator."
   GOAL-BELOW      "Place a suitable container below the oil pan."
   LESS-THAN-BELOW "50 RPM below the maximum speed"
   ```

   Although the preparation of the data can be handled automatically, this last step is presently accomplished by human analysis of the data.

2. *Identify Potentially Ambiguous Role Assignments*.

   (a) First, we list all semantic roles with their structural patterns:

   ```
   PATH-ABOUT    "about NP" "Wrap the wire about the spindle."
   REFERENT      "about NP" "instructions about this procedure"
   VICINITY-OF   "about NP" "about 5 mm"

   LOCATED-ABOVE "above NP" "above the operator station"
   MORE-THAN     "above NP" "above 100 degrees centigrade"
   ```

   (b) Then we notice where the same structural pattern is associated with different semantic roles (as in the cases shown above). These patterns will introduce syntactic ambiguity into the system unless their attachment is semantically restricted. These patterns and sentences which match them are automatically extracted produced as data for the Semantic Role Analyzer (SRA) tool, which is described in Section 7.

3. *Create Mapping Rules*.

   We automatically create all of the potential mapping patterns for each argument/semantic role pair:

```
SEMANTIC ROLE         GRAMMATICAL FUNCTION
-------------         --------------------
PATH-ABOUT            (pp ((root ``about'')))
REFERENT              (pp ((root ``about'')))
VICINITY-OF           (pp ((root ``about'')))
AGENT                 (subject)
AGENT                 (pp ((root ``by'')))
THEME                 (object)
THEME                 (subject)
```

4. *Create Mapping Rule Hierarchy*

   Since many head-argument mappings contain repetitive patterns which are shared among members of a mapping class (for example, a class of verbs which exhibit the same mapping behavior), mapping rules are grouped into classes in a hierarchical structure which eliminates redundancy and speeds the process of knowledge acquisition (for more details, see (Mitamura, 1989) or (Mitamura and Nyberg, 1992)). This step requires analysis by the source language linguist.

### 6.2. Methodology: Target Language Mapping Rules

The target language mapping rules map interlingua structures onto target f-structures. Lexical mapping rules are created automatically from the source and target bilingual lexicon. Since each source lexicon entry contains a pointer to an interlingua head concept (cf. Figure 5), the source and target lexicons can be linked via the interlingua head to produce interlingua-to-target-head lexical mapping rules for each interlingua head and the associated target language lexeme(s).

Argument mapping rules are constructed from the list of semantic roles identified by the source mapping rules and the semantic features present in the interlingua. Once mapping rules have been written which map from semantic roles to syntactic patterns (analogous to those shown in Section 6.1), these mapping rules can also be arranged into a mapping hierarchy. This is accomplished by browsing the target language corpus for examples which fit the target patterns implied by the mapping rules, and associating them with the particular head words which exhibit the mapping behavior in question[4].

### 6.3. Results and Discussion

In the case where there is potentially a one-to-many mapping between interlingua concept and target language lexeme, the KANT lexical selection module plays an important role in producing the appropriate choice (for examples of context-dependent lexical selection, see (Mitamura, Nyberg and Carbonell, 1991)). In a technical domain, the majority of the terms are technical and have a straightforward one-to-one translation; however, general words (e.g., verbs like REMOVE which translates to many words in French) and some technical words (e.g., VALVE, which can be translated different ways in French) exhibit one-to-many lexical choice. Once lexical mappings are produced automatically from the source and target lexicons, the one-to-many mappings are extracted. These undergo a further refinement step where the target language linguist adds contextual patterns to the mapping rules to make the appropriate distinctions for lexical selection (cf. Mitamura, et al., 1991).

[4]The process of target language mapping rule acquisition is currently under active development.

There are also cases where an interlingua head concept maps onto a complex structure in the target language (e.g., *start* → *mettre en marche*). Such cases (where the target language lexical entry is a string rather than a single lexeme) are also extracted automatically and refined by hand.

## 7. Domain Model

Although it is possible to reduce ambiguity by limiting the use of certain kinds of sentence structures, some types of phrases which introduce a high level of ambiguity cannot be ruled out. In English, such phrases include prepositional phrases and noun modifiers. To resolve the ambiguity introduced by multiple possible phrase attachments, KANT uses an explicit domain model to narrow the set of potential interpretations. In conjunction with mapping rule development, which specifies the allowable set of semantic roles and semantic role assignments, construction of the domain model provides the set of semantic role filler restrictions that eliminate spurious attachments.

### 7.1. Methodology

The steps taken to construct a domain model are as follows:

1. *Construction of Concept Frames*.

   An initial set of concept frames is created from the Source Lexicon. A unique concept is associated with each [word, part-of-speech] pair. Any general class concepts created during the assignment of semantic role filler restrictions are added to the set of concepts.

   The data which is extracted from the source corpus during the creation of mapping rules is passed through the Semantic Role Analyzer (SRA) tool. This tool accepts a particular set of sentences and the potential semantic role assignments as input data. For each example sentence, it queries the knowledge engineer about which semantic role assignment is the appropriate one. Then the semantic role fillers are extracted from the examples and used as semantic restrictions in the domain model. For example, if the sentence *Use the lifting eyes on the engine* was encountered during semantic role assignment for on-PPs, then the engineer would select between the ON-LOCATION semantic role attachment to the phrase *lifting eyes* or to the verb *use*. Since the the former choice is the one that makes sense in the domain, the tool would produce a semantic role filler for `*O-ENGINE`:

   ```
   (*O-LIFTING-EYES (ON-LOCATION *O-ENGINE))
   ```

   The semantic role restrictions produced by SRA are automatically merged with the concept definitions in the domain model to restrict the set of possible attachments for each semantic role.

2. *Building a Concept Hierarchy*.

   Once the creation of concept frames and semantic role restrictions is completed, the role restrictions are generalized by the following actions:

   (a) Role restrictors are grouped into classes;

   (b) More general concept frames are created to represent these classes;

   (c) The original role fillers are replaced with the generalized role filler, and inheritance links are created from the specific role fillers to the generalized concept;

(d) At run time, KANT uses a cached inheritance strategy for rapid access to the semantic restrictions.

For example, the PRODUCT-LINE semantic role can be filled by any major product type (tractor, engine, loader, etc.). Automatic extraction of the actual role fillers from the corpus examples would produce a list of all those attested in the corpus, and the lexicographer can then provide the appropriate class description (e.g., *O-PRODUCT-LINE) which can inherit from the original role fillers and any additional concepts which fit into the class. This is crucial to ensure that the system achieves general coverage beyond the examples in the corpus.

### 7.2. Results and Discussion

We have successfully created an initial set of concept frames for the 60,000 lexemes in the Source Lexicon. We are currently in the process of extracting semantic role filler restrictions from this tagged corpus and building a concept hierarchy. These techniques can be of use to any system that requires semantic pattern rules for analysis or generation, even if the system does not contain a separate knowledge base and interlingua representation.

## 8. Conclusion

We have completed the process of source language definition for a large domain of approximately 60,000 words and phrases for heavy equipment manuals. As part of this work, we have successfully applied techniques for automatic source and target lexicon acquisition. We are currently in the process of building source and target mapping rules and a domain model for the same domain using the techniques presented in the previous sections. Although currently applied to English source and French target text, the knowledge acquisition procedures are general enough to be used with any domain-oriented corpus and for any language.

Use of automatic processing in knowledge acquisition has shifted the focus of human effort away from tedious, time-consuming item-by-item knowledge entry. Whenever possible, the system developers work to refine automatically-generated knowledge sources to ensure consistency and coverage. This shift in effort allows MT applications to be constructed in large domains which would otherwise require too much effort. Our ongoing KANT application in the heavy equipment domain demonstrates that a large corpus of 50 megabytes of text can be analyzed automatically to produce knowledge sources (lexicons, grammars, mapping rules, domain model) and also “value-added” resources for human refinement (tagged corpora, KWIC browsers for source/target corpora, aligned source/target contexts, etc.).

Our short-term plans are to complete the process of mapping rule and domain model acquisition for this domain during 1993. Our long-term goals are to first evaluate the knowledge acquisition methodology following completion of the first application, then perform necessary generalizations so that the knowledge acquisition software can be reused for any domain and languages desired, and finally to increase the level of automation further in knowledge source creation.

## Acknowledgements

We would like to thank all the members of the KANT project team, including James Altucher, Kathy Baker, Alex Franz, Susan Holm, Kathi Iannamico, Pamela Jordan, John Leavitt, Daniela and Deryle Lonsdale, and Will Walker. We would also like to thank Claude Doré of Taurus Translations for his help in designing the KANT vocabulary translation tool. We would also like to express our gratitude to our colleagues at Carnegie Group and Caterpillar for their participation in the project.

## References

[1] Carbonell, J., T. Mitamura and E. Nyberg (1992). "The KANT Perspective: A Critique of Pure Transfer (and Pure Interlingua, Pure Statistics, ...)", *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*, Montreal, June 25-27.

[2] Farwell, D., L. Guthrie and Y. Wilks (1992). "The Automatic Creation of Lexical Entries for a Multilingual MT Systems," *Proceedings of COLING 1992*, Nantes, France, July.

[3] Francis, W. and H. Kuĉera (1982). *Frequency Analysis of English Usage*, Boston, MA: Houghton Mifflin.

[4] Goodman, K. and S. Nirenburg (eds.) (1991). *A Case Study in Knowledge-Based Machine Translation*, San Mateo, CA: Morgan Kaufmann.

[5] Grishman, R. and R. Kittredge (eds.) (1986). *Analyzing Language in Restricted Domains: Sublanguage Description and Processing*, Hillsdale, NJ: Lawrence Erlbaum.

[6] Grishman, R. and M. Kosaka (1992). "Combining Rationalist and Empiricist Approaches to Machine Translation," *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*, Montreal, June 25-27.

[7] Grishman, R. and J. Sterling (1992). "Acquisition of Selectional Patterns," *Proceedings of COLING 1992*, Nantes, France, July.

[8] Mitamura, T. (1989). *The Hierarchical Organization of Predicate Frames for Interpretive Mapping in Natural Language Processing*, PhD thesis, University of Pittsburgh.

[9] Mitamura, T., E. Nyberg and J. Carbonell (1991). "An Efficient Interlingua Translation System for Multi-lingual Document Production," *Proceedings of Machine Translation Summit III*, Washington, DC, July 2-4.

[10] Nyberg, E. and T. Mitamura (1992). "The KANT System: Fast, Accurate, High-Quality Translation in Practical Domains," *Proceedings of COLING 1992*, Nantes, France, July.

[11] Sekine, S., S. Ananiadou, J. Carroll and J. Tsujii (1992). "Linguistic Knowledge Generator," *Proceedings of COLING 1992*, Nantes, France, July.

[12] Utsuro, T., Y. Matsumoto and M. Nagao (1992). "Lexical Knowledge Acquisition from Bilingual Corpora," *Proceedings of COLING 1992*, Nantes, France, July.

[13] Zernik, U. (ed.) (1991). *Lexical Acquisition: Exploiting On-line Resources to Build a Lexicon*. Hillsdale, NJ: Lawrence Erlbaum.