Machine Translation, 9:3, 101–133 (1995)


# Acquisition of Large Lexicons for Practical Knowledge-Based MT

DERYLE LONSDALE lonz@cs.cmu.edu
TERUKO MITAMURA teruko@cs.cmu.edu
ERIC NYBERG ehn@cs.cmu.edu
*Center for Machine Translation, Carnegie Mellon University, Pittsburgh, PA 15213*

*Received July 1, 1994; Revised April 15, 1995*

**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 lexical knowledge. 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 magnitude of the development task for 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 viable approach. In this paper we discuss the corpus-based knowledge acquisition methodology used in KANT, a knowledge-based translation system for multilingual document production. This methodology can be generalized beyond the KANT interlingua approach for use with any system that requires similar kinds of knowledge.

**Keywords:** knowledge-based machine translation, conceptual coverage, on-line lexical acquisition, lexical mapping, phrasal substructure

## 1. Introduction

Knowledge-Based Machine Translation (KBMT) is based on the premise that successful machine translation requires the use of explicit levels of linguistic representation (morphology, grammar, semantics, etc.) for domain lexemes. Knowledge is both declarative (e.g., lexical entries, concept definitions) and procedural (e.g., grammar rules, mapping rules, etc.). A *semantic transfer* system uses declarative knowledge and rules to map from the semantics of the source language to the semantics of the target language. An *interlingual* KBMT system makes use of an additional intermediate level of representation, which is intended to be independent from the source and target language semantic representations. It has been shown 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).

### 1.1. KBMT at CMU/CMT

The first KBMT system built at the CMT was a knowledge-based, interlingual system for doctor-patient communication in English, Japanese and German (Tomita et al., 1987). This system was the first working prototype of the basic KBMT/KANT-type architecture combining analysis and generation based on unification grammars, with semantic interpretation and mapping of interlingua structures.

The first application of this prototype for document translation was the KBMT-89 system, which has been described in detail in a special issue of *Computers and Translation* (volume 4, 1989). KBMT-89 performed bi-directional English-Japanese translation of IBM PC installation manuals. The research goals of the KBMT-89 project were fulfilled by the demonstration in early 1989 of an integrated run-time system which produced high-quality output. Nevertheless, there were two considerations which limited the practical potential of the system:

- *Ambiguity in the Source Text.* Although KBMT-89 included a dialog management component which communicated with the user whenever the source text was ambiguous (Brown, 1991), it was possible for the user to write sentences that were sufficiently complicated and ambiguous that system performance was reduced to unacceptable levels.

- *Run-Time Performance.* Although the LR Parser Compiler used in the KBMT-89 system supports compilation of unification grammars into very fast run-time parsers, much of the other knowledge used in KBMT-89 was not compiled into an optimally efficient run-time form. The run-time system required 1-2 minutes to process a single sentence, which is fine for a demonstration but too slow for practical applications demanding high throughput.

### 1.2. The KANT System

The KANT system, also developed at CMU, is a direct descendent of KBMT-89 and incorporates a similar software architecture. Nevertheless, KANT is designed to work on a much narrower range of translation problems than KBMT-89, with the resulting advantage of improved accuracy and better run-time performance. The primary characteristics of KANT include:

- *Focused Domain.* KANT is designed to work in a controlled, technical domain. A typical KANT application is translation of technical documents concerned with some process or product (e.g., electrical utility management (ESTRATO), heavy machinery documentation (CATALYST). This makes it feasible to construct a semantic model of the application domain for use during translation.

- *Controlled Source Language.* KANT requires that the source vocabulary and grammar used by the application be controlled. The vocabulary is limited to just those word/meaning pairs that are necessary for the domain. The source grammar is limited to just those syntactic constructions that are required to productively author text in the domain. This limits the complexity of the source text, reducing ambiguity and improving accuracy and run-time performance.

- *Large Scale.* KANT is intended for practical applications, so it involves scaling up lexicons and rule bases to a large degree. For example, in the CATALYST application, there are approximately 60,000 domain concepts.

- *Interlingua.* KANT makes use of a detailed interlingua representation for the meaning of each source sentence (Lonsdale, Franz, and Leavitt, 1994). This supports modular system design (since there are no pair-wise dependencies between source and target languages) and supports accurate translation by capturing the necessary semantic information.

- *Multiple Target Languages.* KANT is designed for use with multiple target languages, thus making full use of the modular potential of interlingual MT systems.

- *Software Architecture.* The KANT architecture consists of software that remains constant from application to application, coupled with lexicons, a domain model, grammars, and mapping rules for a particular application. This decoupling of code and knowledge makes it easy to extend the system to new languages or domains, or to reuse any part of the system for a new domain. This also supports pre-compilation of declarative knowledge structures into a faster run-time form, improving system performance (Mitamura, Nyberg, and Carbonell, 1991). This type of software design also enhances long-term maintainability of finished applications as new processes or products are added to the domain.

### 1.3. How KANT Works

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;

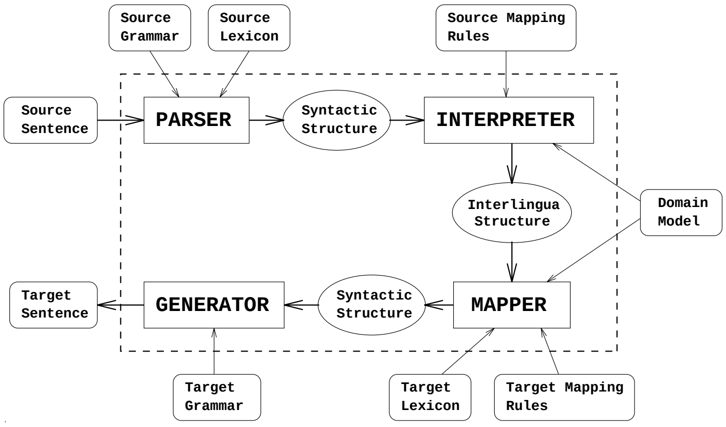


*Figure 1.* The Run-Time Architecture of KANT

- 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.

### 1.4. KANT Status

The largest KANT application to date is the CATALYST system, an ongoing development for Caterpillar, Inc. The system translates documents written in Caterpillar Technical English (CTE). The domain of Caterpillar product literature is quite large, requiring many document types and a technical vocabulary of about 60,000 words and phrases. These documents are to be translated to 11 different target languages; the first module, French, has been completed and delivered to Caterpillar. During 1993 and 1994, the CATALYST French system underwent an intensive test/debug/evaluation cycle which improved the system's performance before delivery (Nyberg, Mitamura, and Carbonell, 1994). The Spanish target language will be completed in 1995, and German in 1996. Development of Italian and Portuguese are underway.

### 1.5. The Problem of Large-Scale Lexicon Acquisition

During the development of KANT applications, we have tried whenever possible to make use of on-line resources (dictionary, corpora, etc.) and analysis techniques (indexing, alignment, extraction, etc.) to support the process of acquiring domain and linguistic knowledge for an application (Mitamura, Nyberg, and Carbonell, 1993; Leavitt et al., 1994). Given the apparent success of the CATALYST system, we can say that KBMT systems can indeed be scaled up, and that corpus analysis and on-line resources are certainly useful. But for highly-accurate technical translation, complete automation does not seem to be a feasible goal because the available resources are often insufficient.[1] In the remainder of this paper, we present the knowledge acquisition methods that we used during lexicon development for English and French in CATALYST. In addition to the description of the methods and results, we try wherever possible to mention particular challenges, tools, and lessons learned.

### 1.6. Lexicon Acquisition in KANT

Because the focus of this paper is on lexical 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).

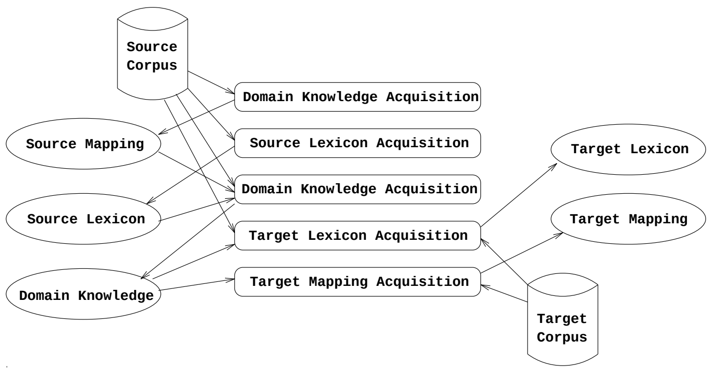


*Figure 2.* Knowledge Acquisition in KANT

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 this paper, we will address the following aspects of lexicon acquisition in KANT, focusing on English as a source and French as a target:

- Conceptual Coverage of the Vocabulary
- Source Language Lexicon
- Source Lexical Mapping
- Target Lexical Mapping
- Target Language Lexicon

## 2. Conceptual Coverage of the Vocabulary

Whenever a company documents its output, whether in manufactured goods, services, or other types of product offerings, it will necessarily use words and phrases which reflect the unique nature of that company's contributions to the economic landscape. A company-specific vocabulary evolves along with the company itself, and often reflects the corporate philosophy, culture, and image. In many cases it is proprietary, and is sometimes even legally protected. Hence the inventory of words and phrases that a company uses to present itself to customers and others is an important asset. When documentation must be translated into another language, a cross-linguistic dimension is added to these considerations.

The KANT approach includes, like any large-scale MT enterprise must (Galinski, 1988), a terminological component designed and implemented to address the complex vocabulary-related problems that are associated with translation. In this section we survey how vocabulary was assessed, collected, processed, refined, and chunked during the development of a large-scale machine translation system.

### 2.1. Vocabulary Types

For our purposes we will distinguish between two basic types of vocabulary: general (or non-technical) vocabulary, and specialized (or technical) vocabulary. This distinction can also be considered across another dimension: function words versus content words.

For English, some general vocabulary includes (usually commonly occurring) function words such as prepositions (e.g., *in*, *under*, *of*), conjunctions (e.g., *or*, *if*, *whether*), determiners (e.g., *our*, *the*, *any*), pronouns (e.g., *it*, *they*), and auxiliary verbs (e.g., *may*, *can*). These words express basic relationships and connections between objects and ideas, and belong to closed (i.e., non-productive) classes. Sometimes these words are strung together (e.g., *out from*, *any of*), though such collocations still retain a non-technical aspect of usage.

Unlike function words, the class of content words is highly productive. Content words express the objects, properties, actions, and manners which are connected by function words. We may make a further distinction between general (non-technical) content words, and specialized (technical) ones. The former would include the most commonly occurring nouns (e.g., *machine*, *work*, *oil*, *air*), verbs (e.g., *see*, *try*, *go*, *stop*), adverbs (e.g., *quickly*, *never*, *now*), and adjectives (e.g., *big*, *fast*, *difficult*). It should be noted, however, that apparently general content words are often used by analogy, metaphor, etc., to express ideas of a technical nature.

The other type of content words are specialized or technical content words. These constitute a much more open-ended class of vocabulary. Technical documentation is full of this type of lexical item, which may include sophisticated technology-based terms for product names, part names, documentation titles, form numbers, acronyms, and measurements. Even specialized verbs, adjectives, adverbs can emerge with new technological developments, and thus enter a language's store of content words.

### 2.2. Conceptual Coverage

As discussed earlier, the KANT system is designed to perform translation of texts of a certain subject area, called the domain. Since it is an interlingua-based system, care must be taken to ensure that anything expressed in the vocabulary of the entire domain can be likewise expressed by an interlingua.

A methodological question arises on this point: how does one ensure that an interlingua is sufficiently powerful and unambiguous to adequately represent an entire domain of application? What procedures must be followed to build such a representation? Tsujii (1988) investigates these issues and proposes three general ways to approach the task of choosing and implementing a domain-comprehensive interlingua. One could:

- consider the domain and enumerate *a priori* the concepts, processes, and relationships required for its treatment (i.e., a top-down approach)
- consider the (disambiguated) lexical content expressed by text discussing the domain, and then define appropriate sets, hierarchies, and other relevant relationships (i.e., a bottom-up approach)
- (re-)express all relevant aspects of the domain with respect to a highly restricted set of semantic primitives (i.e., a decompositional approach)

We have found that a combination of the first two approaches works to ensure complete lexical coverage of a large-scale concept inventory.

Our top-down effort involved establishing a circumscription of the domain addressed, discourse styles, and typical document structure. This led to an inventory of discourse entities including machines, persons, environmental conditions, occupations, organizations, measurements, complex operations, etc. Valuable insight was provided by domain experts who provided experience and introspection as we considered these questions. This rationalist approach, while costly because of its involvement of domain experts, is probably unavoidable given the background knowledge required.

At the same time, we carried out extensive bottom-up identification of the domain through extraction of knowledge by automated corpus analysis techniques. A top-level discussion of this process will be given in the following section on the source lexicon. This empiricist approach, based on large-scale data analysis, ensured completeness of coverage.

### 2.3. Standardization

As the vocabulary used in domain-specific documents was processed through corpus analysis techniques, an identifiable subset of English emerged. It was clear, however, that a certain amount of canonicalization of vocabulary usage was possible, even within this domain. In particular, the following steps were carried out:

- orthography was normalized; American standard spelling was adopted, variants were eliminated, errors were discarded:

  *exhaust gasses = exhaust gases*
  *9 tooth dog clutch = nine-tooth dog clutch*

- reduced forms were canonicalized; abbreviations and acronyms were judged for comprehension and informativeness, measurement unit abbreviations were made consistent:

  *digital lcd display = digital liquid crystal display*
  *module s/n = module serial number*

- word-separation was standardized; the use of hyphen, slash, and space between terms was dovetailed:

  *air to air aftercooler = air-to-air aftercooler*
  *blowby/airflow instrument = blowby/air flow instrument*
  *air/fuel mixture = air-fuel mixture*

- nominal compounding was streamlined; morphological variants were merged, equivalent word orderings were collapsed:

*vibratory motor = vibrator motor = vibration motor*
*air condition system = air conditioner system = air conditioning system*
*windshield washer/wiper = windshield wiper/washer*

This standardization effort reduced the stock of source-language terms by several thousand tokens. Besides providing a cleaner, more consistent source text vocabulary, this step also simplified the development of all related system components. The end result was a high-quality domain terminology based on the vocabulary used to redact hundreds of documents.

### 2.4. Lexical Chunking

Once the source-language terms were identified, a decision had to be made regarding the granularity of lexical treatment by the system. We have found that rather than attempting to attain the lowest-level (i.e., sublexical) grain size, it is quite useful to perform lexical chunking. We thus have chosen to consider as atomic such highly lexicalized items as fixed phrases, technical nomenclature, and company-specific terminology. In our mapping of lexical items to concepts, we created unitary concepts from collocations of primarily content words, wherever possible. Idiomatic and other fixed-form lexical collocations were also considered unitary. We therefore have chosen to create concepts like *O-KNURLED-FULL-ZONE, *A-BREAK-APART, *P-AIR-TO-AIR-AFTERCOOLED, *M-BY-ACCIDENT, and *U-JOULE, rather than to decompose these any further.[2] Of course, straightforward modification by (single-word and phrasal) adjectives and adverbs was taken to be productive, thereby reducing the total concept inventory.

Our grain-size was set based on an examination of the communication content of the documents we address. Resisting the tendency, given the complexity of natural language phenomena, to favor the creation of complicated interlingua structures, we based our work on practical notions of lexical chunking. Given the constraints on the input text, and the limited domain of application, we were able to dispense with the complex attitudinal, situational, and discourse-related aspects of sub-lexical description seen in other interlingua systems. Our goal was to design a minimalist representation, considering both the breadth and depth of the domain addressed, and avoiding the opposing pitfalls of over-complexity and under-specificity.

### 2.5. Structural Aspects

Though the source and target languages express basic relations via function words as described above, in an interlingua system these should be abstracted away in favor of a more semantic-primitive approach. For example, prepositions are represented, as explained in the next chapter, by primitive slots (not concepts) like **PATH-ABOUT**. Since this type of abstraction does not introduce a strictly lexicalist mapping in the target language, but rather a structural one, it falls beyond the scope of this paper. We will therefore not discuss the mapping of function words for target-language mapping. but rather focus on content words, their instantiation as concepts, and their mapping to target language expressions.

## 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 is used to extract candidate vocabulary entries, which is then subjected to human refinement. For the CATALYST application, we used a corpus of about 53Mb of Caterpillar on-line text files, including different products and document types.

### 3.1. Creation of Source Language Lexicon

The steps taken to construct a lexicon from the Raw 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 and Kučera, 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[3].
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.

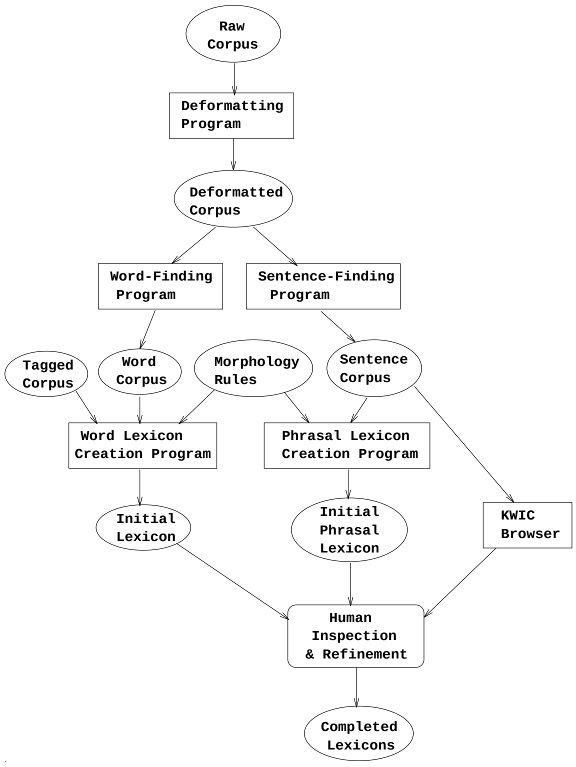


*Figure 3.* Automation of Lexicon Acquisition

```
((: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 4.* Example Lexicon Entry

An example of a finished lexicon entry is shown in Figure 4. The `:ROOT`, `:POS`, `:CONCEPT`, `:SYL-DOUBLE`, and `:SYN-FEATURES` fields are created automatically with default values; the `:FREQUENCY` field contains values for the occurrence count of the word in the source corpus in single-word and phrasal contexts. 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

As mentioned above, this method was applied to a large corpus of Caterpillar documents (approximately 53 megabytes of deformatted text). A single-word lexicon of about 9000 items was extracted from the corpus. A phrasal lexicon of about 51,000 phrases was also extracted from the corpus.

The Tagged Corpus contains some part of speech assignments which are not part of the application domain.[4] 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 5. 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 be 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.

```
(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 5.* Example Phrase Refinements

### 3.3. Discussion

It is arguably more efficient 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 manually[5]. 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 total effort in human refinement of lexical entries. We would also like to investigate the potential use of automatic tagging 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. Source Lexical Mapping

In KANT, source language sentences are syntactically analyzed and semantically interpreted, and the resulting interlingua representation is used as an intermediate stage in multilingual translation. A set of mapping rules is used to map source language lexical items onto interlingua structures.

There are two types of source lexical mapping rules used in KANT: Lexical Mapping rules and Argument Mapping rules (Mitamura, 1989). Lexical Mapping rules map a content 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).

The majority of general content words are limited in the lexicon to one sense per part of speech. However, there are a few hundred terms which have more than one sense per part of speech. In this case, each sense maps to a different concept node. Basically, each content word with a unique sense maps to a unique concept. Phrases are defined with a single sense and map to a single concept (e.g., *O-SUMMING-VALVE). Some function words, such as conjunctions, determiners, and auxiliary verbs, are mapped onto a feature-value pair. For example, "the" is mapped onto the feature-value pair `(reference definite)`.

Argument Mapping rules map a grammatical function onto a semantic role. 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.

### 4.1. Source Language Argument Mapping Rules

The construction of source language mapping rules requires 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.*

In order to identify the full set of semantic roles necessary for representing grammatical functions in the application domain, the corpus is searched automatically for source language constructions which are associated with general semantic roles (e.g., "with" + NP $\rightarrow$ INSTRUMENT). The corpus examples are then interactively presented to an expert who indicates which of the possible semantic role assignments for these constructions are actually indicated by domain use.

(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.

```
Do not allow pressure to drop below 2750 kpa.
Reinforce the flange below the slot.
```

(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.*

Once the set of allowable semantic roles has been narrowed to just those indicated by the domain examples, then attention is focused on those role assignments which are still potentially ambiguous (i.e, there is a single grammatical function which has more than one possible semantic role assignment). These are automatically extracted from the set of semantic roles by searching for those with identical contexts.

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

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

LOCATED-ABOVE   ``above NP'' "above the operator station"
MORE-THAN-ABOVE ``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 and produced as data for 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 LOCATED-ON 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 (LOCATED-ON *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.

3. *Create Mapping Rules.*

   For the set of ambiguous semantic role assignments, we create restrictive mapping rules which license syntactic attachment based on the semantics of the phrase head and the potential semantic role filler. The rules are created automatically. Then the original set of corpus examples extracted for the ambiguous contexts is used to test the 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-ABOUT       (pp ((root ``about'')))
VICINITY-ABOUT       (pp ((root ``about'')))
AGENT                (subject)
AGENT-BY             (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.

### 4.2. Discussion

Lexical mapping rules are automatically generated from the domain lexicon, which contains pointers to concept nodes. Therefore, the total number of all lexical mapping rules is proportional to the number of lexical entries, with each sense having a separate entry in the lexicon.

Our semantic roles for PP consist of a semantic role followed by a preposition. Therefore, PPs which have the same semantic role but different prepositions are represented in different ways. For example, `LOCATED-IN` indicates the semantic role `LOCATED` and the preposition "in", whereas `LOCATED-AT` indicates the same semantic role expressed by the preposition "at". When we develop different types of semantic roles, we face a question of granularity of roles. If the roles do not make enough distinctions for accurate generation, the quality of translation suffers (for example, if no distinction is made between `LOCATED-IN` and `LOCATED-AT`). On the other hand, if we make semantic distinctions which are unnecessary for any target language, then effort may be wasted. Therefore, we need to attain a level of granularity which is constrained yet expressive enough for accurate translation.

## 5. Target Lexical Mapping

The result of the source mapping process as described in the previous section is an interlingua representation of the input source-language text. In this section we address target lexical mapping, the process whereby portions of the interlingua are reexpressed in the vocabulary of the target language. In part, this stage of mapping involves processing of individual concepts in isolation, since each concept is usually mapped onto a target term. Often, though, lexical mapping also relates several concepts, and mediates between realizations based on these relationships. In this section we discuss and illustrate the various lexical mapping types. Of course, there are also structural mapping processes involving non-lexical aspects, but these would lead us beyond the scope of this paper.

### 5.1. Overall System Context

The mapper is a component of the system which recursively traverses the interlingua, stopping at each level to examine slots and their fillers (features, concepts, and nested interlinguas). Testing a hierarchy of rule declarations, the mapper performs a structure-building operation called mapping. The goal and result of mapping is a target-language f-structure whose contents reflect the properties of the interlingua, expressed in terms of the syntactic and lexical properties of the target language. By this process concepts receive a lexical realization, and structural relationships expressed elsewhere in the interlingua are generally realized as function words.

In the following discussion we will survey the different types of concept-lexical mappings. Various types of mapping rule declarations will be explained, and control decisions about selection will be mentioned. The process of setting up equivalences between concepts and target language terms will also be discussed. Finally, further research questions are addressed.

### 5.2. One-to-One Encodings

In the ideal situation, all translation mappings would require a simple one-to-one mapping between concept and target lexical item. A simple declaration like the following suffices for establishing the correspondence between the two in such cases:

```
("en fibre de verre" :PARENTS (ADJ) :ENCODES (*P-FIBERGLASS))
```

This declaration indicates that the property concept `*P-FIBERGLASS` is realized in the target language as the term *en fibre de verre*. Every instance of this concept, whether predicative or attributive, will receive the declared realization. Other such examples include:

```
("axe de trémie" :PARENTS (NOUN) :ENCODES (*O-HOPPER-PIN))
("déneigement" :PARENTS (NOUN) :ENCODES (*O-SNOW-REMOVAL))
("limiteur de surtension" :PARENTS (NOUN) :ENCODES (*O-PREREGULATOR))
```

Here a target language noun phrase realizes an object concept such as `*O-HOPPER-PIN`. In each case, no other realization will occur for the relevant concept, and no other concept will receive the declared realization. Note that a concept may be realized as a single word or as a multi-word phrase.

Naturally, such a simplistic approach does not meet the requirements for true translation, where correspondences are not as direct. This requires other types of declarations setting up more complex mappings.

### 5.3. Many-to-One Encodings

It is unavoidable that some concepts will be realized in the target language synonymously, given a reasonable level of granularity. Thus, we must allow for many-to-one realizations, declared in this fashion:

```
("monocorps" :PARENTS (ADJ)
            :ENCODES (*P-SINGLE-SECTION *P-SINGLE-BARREL *P-ONE-SECTION))
```

Here we have an example of a target language term which is the realization of three different (but largely synonymous) concepts from the domain. The adjectival term *monocorps* completely spans the only meaning expressed by all three of these concepts. A few other such mapping declarations follow:

```
("peut-être" :PARENTS (ADV) :ENCODES (*M-MAYBE *M-PERHAPS))
("proportionnellement" :PARENTS (ADV) :ENCODES (*M-PROPORTIONALLY
                                                *M-PROPORTIONATELY))
("amortir" :PARENTS (VERB) :ENCODES (*A-DEADEN *A-DAMPEN *A-CUSHION))
("détendre" :PARENTS (VERB) :ENCODES (*A-SLACKEN *A-RELIEVE *A-RELAX))
("tourbillonner" :PARENTS (VERB) :ENCODES (*A-WHIRL *A-SWIRL))
("latéral" :PARENTS (ADJ) :ENCODES (*P-SIDE-TO-SIDE *P-SIDEWAYS *P-LATERAL))
("à trois tiroirs" :PARENTS (ADJ) :ENCODES (*P-THREE-SPOOL *P-THREE-STEM))
```

### *5.3.1. One-to-many encodings*

Concept-to-lexical mapping can also be a one-to-many process. If a concept is realizable by more than one target-language lexical item, the alternatives can be declared with selection probabilities if so desired. These selection probabilities can be set, for example, based on a frequency analysis of the occurrences of each possibility in the target language corpus. Below are a few examples:

```
(select (0.3 (rule *O-FUEL-ROD :LEX "biellette d'injection"))
        (0.7 (rule *O-FUEL-ROD :LEX "tige d'injection")))
(select (0.2 (rule *O-BUMPER-SENSOR :LEX "capteur sur le pare-chocs"))
        (0.8 (rule *O-BUMPER-SENSOR :LEX "capteur du pare-chocs")))
(select (0.4 (rule *O-BACKHOE :LEX "rétrocaveuse"))
        (0.1 (rule *O-BACKHOE :LEX "rétro"))
        (0.5 (rule *O-BACKHOE :LEX "pelle rétro")))
(select (0.5 (rule *P-REDUCED :LEX "réduit"))
        (0.5 (rule *P-REDUCED :LEX "diminué")))
(select (0.5 (rule *P-VERSATILE :LEX "polyvalent"))
        (0.5 (rule *P-VERSATILE :LEX "universel")))
(select (0.5 (rule *M-INADVERTENTLY :LEX "par mégarde"))
        (0.5 (rule *M-INADVERTENTLY :LEX "par inadvertance")))
(select (0.1 (rule *A-REDIRECT :LEX "rediriger"))
        (0.9 (rule *A-REDIRECT :LEX "réacheminer")))
```

Of course, these rules could be replaced by one-to-one rules, but rules like these, if carefully crafted, contribute to the generation of target texts with greater lexical variety.

## 5.4. Lexical Selection

In some cases, the granularity of the concept in the domain is not fine enough to reflect distinctions present in the possible target language realizations. In such cases, it is necessary to write lexical selection rules which are based on contextual properties present in the interlingua. For example, the concept `*A-OPERATE` can be rendered in French by (at least) the two translations *conduire* or *actionner*, depending on whether the theme is a driveable machine or a movable part (i.e., *Operate the tractor.* versus *Operate the lever.* respectively).

Lexical selection declarations like the following can ensure the proper mapping:

```
(rule *A-OPERATE
      :test (:sem (theme DRIVEABLE-MACHINE))
      :lex "conduire")
(rule *A-OPERATE
      :test (:sem (theme MOVEABLE-PART))
      :lex "actionner")
```

In these lexical mapping rules, the callouts `(theme DRIVEABLE-MACHINE)` and `(theme MOVEABLE-PART)` test the concept in the theme slot of the frame in question, and evaluates whether the concept is subsumed under the relevant concept class in the domain model.

Next we turn to another example of context-induced lexical selection for concept realization. Concepts may be realized in full-form lexical items, or sometimes as acronyms, initialisms, or abbreviated forms. Reduced forms are signalled by a feature like `(abbreviation +)` in the interlingua, associated with the concept:

```
(rule *U-ATMOSPHERE
      :test (:sem (abbreviation +))
      :lex "atm")
(rule *U-ATMOSPHERE
      :test (:sem (:not (abbreviation +)))
      :lex "atmosphère")
```

Here we have the pressure measurement unit *atmosphere* being translated as either a full-form or reduced-form target term, based on the presence of a feature-value pair in the interlingua, reflecting source-language usage.

### 5.5. Prototyping and Scaling up

In the prototype system, target lexical mapping was set up by hand, based on specifications for the relevant data types. The relevant concepts in the prototype text were first identified; their target language realizations were then determined and verified by consultation with an expert translator for this information. Next, the encodings were recorded in declarative form. Only the most basic patterns of lexical mappings were exemplified in the prototype system.

In order to ensure complete domain coverage for the full production system in terms of lexical mapping, a correspondence between the inventory of the domain concepts and the target lexical domain was required. In addition, each concept had to be assigned one or more target lexical realizations. Given the large inventory of concepts, and the open-endedness of the possible target language expressions, the task involved was substantial. Different types of resources were used; they are briefly discussed in this section.

### 5.5.1. *Lexical resources*

We profited as much as possible during the process of concept-lexicalization assignment from available bilingual resources.

For example, the customer was able to supply us with a bilingual parts database which was used for non-translation-related purposes. In addition, a basic customer in-house bilingual lexicon of technical words and terms, used by the company's translators, was also made available to us. Both required significant reworking and massaging for computational purposes, but the end result justified the work involved.

In addition, we were able to use terms from a few off-the-shelf bilingual dictionaries in machine-readable form. These terms recovered from this source were necessarily of mostly non-technical nature, but complemented the terms from the customer-supplied resources just mentioned.

All of the target terms retrieved from this process were filtered against the target corpus to remove any terminology not deemed current or valid. Some 6,500 usable target terms were recovered from these lexical resources, and were thereby associated with source terms identified with concept names.

### 5.5.2. *Source-target alignment*

Our next step involved recovery of source-target terminology correspondences via bilingual alignment of previously translated documents. Though this target corpus of translations only covered about 11% of the source corpus, it still constituted a valuable resource of the customer's complicated bilingual terminology history and evolution. By processing the target corpus and aligning it with the source corpus, a rich set of translation equivalents emerged.

Our first step was to deformat the target texts. We were able to use the deformatting system developed for the source text analysis, though modifications had to be made to handle the different standards used to encode the 8-bit accented characters found in the French target corpus. The result was a standardized 1,750,000-word corpus of previously translated documents.

After the target corpus was deformatted, a KWIC index was generated for all the single words used in the corpus. This allowed us to browse any word in the corpus, with variable context. Some 19,000 single word types were recovered from the corpus, including several technical terms, measurement units, abbreviations, acronyms, and other special-purpose words.

Next, a nominal compound KWIC index was generated. It features several different types of compositional and collocations, and likewise served as a reference for context-anchored browsing. Over 160,000 complex nominal types were mined from the corpus.

Once the target terms were identified and indexed, an alignment of the source and target terms was calculated. Because of rather wide divergences between the two texts, a dynamic-programming approach was abandoned for a method based on similarity matching between the languages. This was highly successful for us because of the frequent appearance of measurements, numbers, alphanumerics (such as part names), and tagged extralinguistic lexemes which did not change during translation.

A bilingual browser, BiKWIC, was developed to permit easy in-context access to any single-word or multi-word term and its translation(s). This tool, when fed a list of source terms, was used by bilingual non-experts to capture (via a mouse operation) the relevant translation(s) as encountered in the aligned corpora. In this manner translations were found for over 8,000 nominal compounds, and over 4,000 single words which had not yet been recovered from the lexical resources mentioned above. These translations were almost always of highly technical nature, and consultation of a domain-expert translator would have increased costs dramatically.

Once the translations have been collected, the source-term/concept mappings are calculated, and then the concept/target-term mappings are established. For more information about the bilingual alignment and extraction processes, along with screen dumps of the BiKWIC browser, see (Lonsdale, 1994).

### *5.5.3. VTE*

The Vocabulary Translation Editor (VTE) tool was developed to assist in the daunting task of assigning target language terms for all concepts not assigned realizations by the methods mentioned above. This tool was specially designed for use by domain-expert translators, whose time-critical but indispensable participation must be facilitated with user-friendly and task-specific functionality.

The translator must be able to evaluate each concept, browse in context its possible usages, consult the source mapping lexicon for predefined usage constraints, and input one or more target-language lexical realizations for that concept. Wherever possible, the system will aid in positing draft translations (to minimize input keystrokes) based on partial composition patterns already found in related subterms or superterms. As the translator reviews translations supplied by lexical resources or alignment, enters new translations, or accepts draft ones, the target translations are stored into a concept mapping database. Specially designed programs are run against this database to automatically create the different kinds of lexical mapping declarations described in this section.

For a more detailed discussion of the VTE tool and illustrations of its screen formats, see (Leavitt et al., 1994).

#### 5.5.4. *Final result*

In this section we have summarized the process required to scale up the target lexical mapping component of the KANT system. The mixture of corpus analysis tools and methods, as well as specifically designed editors for expert translators, has allowed us to build a large-scale interlingua-based translation system from available corpora and lexical resources. Each target language module to be developed in the future will be based on some combination of these (and newer) methods, although the availability of a target corpus and bilingual lexical resources will vary widely from language to language. The insights gained so far should help us adjust to these uncertainties.

### 5.6. Discussion

In this section we conclude with a few remarks about problems we encountered, and suggestions for future improvements.

First of all, it should be mentioned that the development of target mapping rules requires a unique kind of specialization: a thorough knowledge of translation-related vocabulary issues, the interlingua structure, the target-language generation component, the corpus analysis tools, and even some familiarity with the domain.

The primary lexical resources at our disposal for bilingual term matching, while helpful, were still quite modest: a great deal of expert-supplied work could have been avoided with a more complete lexicon and bilingual corpus. We could have also invested more effort in extracting from the corpus the types of information we often had to pose to domain experts. Furthermore, our alignment technique, while satisfactory for our purposes, could have been applied to a more automatic extraction of term equivalents. We could have introduced automatic stemming techniques into the source and target corpora. A more modular framework (e.g., LEX instead of C) could have been used to simplify the specification of phrase-indexing routines for source and target languages. In all of these aspects, though, we reached what felt like a point of diminishing returns, given the size of the corpora at our disposal and time constraints. With a fuller resource base, we would probably have been tempted to leverage the corpora to a larger degree.

Mention should be made here of the fact that lexical selection is not only lexically driven; discourse factors need to be considered as well. For example, bilingual browsing on the word "heat" revealed that it was translated sometimes as *chauffer*, and sometimes by *réchauffer* and even *préchauffer*. In such cases, the human translator was able, due to discourse context and domain knowledge, to insert these morphologically-signalled bits of extra knowledge that were not explicit in the source text. The KANT system, because of its lack of discourse-referent tracking and fine-grained situational context analysis, is not able to perform such lexical selections. This is an area where the considerations of practicality are adhered to at some expense. One strategy we followed where possible was to encode the most general translation which could properly realize all of the near-synonymous source terms. Of course, use in the source text of the verbs *preheat* and *reheat*, which do exist in the controlled source language, would be more precise in such cases and would always lead to the correct translations. This tension between clarity of source text authoring and directness of translation is obviously one to be sounded out thoroughly in the development of a practical system like KANT.

Finally, in an enterprise as extensive as the KANT system, we keenly felt the tradeoff between research potential and development requirements. Though we were often tempted to bootstrap our efforts with aggressive automation of our work using still-evolving techniques, we sometimes had to satisfy the fixed timetables of our client by resorting to existing but more human-intensive methods. We thus had to strike a balance between state-of-the-art research and time-proven techniques.

## 6. Target Language Lexicon

The target lexicon includes the definitions necessary to realize a given target-language term. This includes all levels of description required by the generation system. In this section we will sample the purpose and structure of the target lexicon, its evolution, and future issues.

### 6.1. Description

In order to generate a target text, a generation system must have a thorough representation of the semantic, structural, and lexical information to be incorporated in the output. Since the output is in graphemic form, even higher-level issues like orthographic adjustments must also be addressed. In this section we investigate the various levels of representation included in the target lexicon, and their motivation.

Syntactic information is the level of representation most explicitly represented in the target lexicon. In this respect the lexicon associates with each target term and phrase an f-structure (or "fs"). As the words are combined into sentences, the respective fs's are combined together in a structural mapping process. At the sentential level, identification of syntactic information such as the major constituents (subject, object, phrases, clauses, conjoined structures, etc.) is necessary for generation. More relevant to us, though, is the granularity of structures represented in the target lexicon, which includes words, nominal compounds, and verbal idioms.

The simplest target lexicon encodings are single words, whose fs usually involves only the major syntactic category and any other lexically-related features. For example,

```
(NOM-F "gâchette" ((ROOT "gâchette")))
```

is a target lexicon entry defining the word *gâchette* as a feminine noun. Because of the template NOM-F, certain feature-value pairs will be added to the defined fs when the lexicon is loaded at run-time; the final result will be the word-specific fs:

```
((CAT noun)(ROOT "gâchette")
 (AGR ((GENDER f)(NUMBER sg)(PERSON 3))))
```

This fs is all the information required for the generation of the target word in question. There is obviously little structural information that needs to be expressed at this level. The target lexicon has some 6500 single-word declarations.

On the other hand, complex structural information is required for multi-word target terms, which may include complex nominals, adjectivals (with prepositional or phrasal complements), adverbials, and even verb idioms. Note the following (rather simple) example, the translation for "main electrical system reset":

```
(PHR "réarmement du circuit électrique principal"
 ((CAT noun) (ROOT réarmement)
  (AGR ((PERSON 3) (NUMBER sg) (GENDER m)))
  (PP
    ((P-OBJ
       ((ROOT circuit) (CAT noun)
        (MODIFIER
           ((ROOT conj-adj)(CAT adj)
             (MEMBER
              (*MULTIPLE* ((CAT adj) (ROOT électrique))
                          ((CAT adj) (ROOT principal))))
             (CONJ ((ROOT "")))))
        (DET ((CAT art) (ROOT le)))
        (AGR ((GENDER m) (NUMBER sg) (PERSON 3))))))
     (PREP ((ROOT de) (CAT prep))))))
```

Here we see a phrasal definition having a head noun (with its lexical features) accompanied by a prepositional phrase modifier consisting of the preposition *de* and PP object having two single-word adjectival postmodifiers. Such nesting becomes even more complex in fs definitions for terms like *sélecteur de commande de profil transversal supérieur gauche*. Several tens of thousands of lexical items involve this type of structure and complexity.

Next we show an example of a verbal lexicon declaration, the translation for the verb "hand-tighten":

```
(VERB "serrer à la main"
  ((CAT verb) (ROOT serrer)
   (PP ((P-OBJ
         ((ROOT main) (CAT noun)
          (AGR ((PERSON 3) (NUMBER sg) (GENDER f)))
          (DET ((CAT art) (ROOT le)))))
        (PREP ((ROOT à) (CAT prep)))))))
```

Here we have a structure headed by a verb which is followed by a prepositional phrase (with preposition and object specified). Several hundred verbal, adjectival, and adverbial idioms are contained in the lexicon with pertinent structural information.

Of course, morphological information is also required by the system and thus is sometimes present in the target lexicon. Earlier mention has been made of the morphological processor which contains procedural rules for performing inflection. Our intent was to make the morphology component as powerful as possible, and thus minimize the amount of explicit morphological information in the target lexicon itself. Thus, word forms are typically represented in the fs in their uninflected, or base forms, along with annotations like `(number plural)` or `(gender feminine)` when morphological inflection is required.

Some phonological information is contained in the target lexicon. For example in French, rules for performing article reduction are in part phonologically based. While morphophonemic rules can account for most of the cases in question, a rote inclusion of some forms must be made. For example, the aspirate/unaspirate distinction is relevant: in *sélecteur de haute gamme*, the word *haute* is aspirated, so the article liaison is blocked; in *soupape d'huile* the word *huile* is not aspirate, and the article is liaised. This holds not only for words beginning with the letter "h", but also for abbreviations, initialisms, and acronyms depending on their representations and pronunciations. This information must be captured in the target lexicon. It can be found in most dictionaries, and was recoverable from the on-line resources we used. Of course, for client-specific terminology not found in any dictionaries, we had to do this manually; the on-line target corpus was useful for this task.

## 6.2. Prototyping and Scaling up

As with most other aspects of the system, the prototype target lexicon was crafted by hand. All target terms which could potentially be generated by the system were identified, analyzed for compositionality, and encoded by hand. Lemmatization was performed for those terms to be treated by the morphology, and f-structures were built by hand for the terms in question. Some 300 terms were encoded in this fashion.

Scaling up the target lexicon has proven to be a time-consuming and labor-intensive task. Fortunately, we have been able to leverage several of the tools and methods from the overall KANT framework to help reduce the complexity of the several tasks. In this section we will discuss this effort.

### *6.2.1. Vocabulary analysis*

We first had to acquire a working intuition for the nature of the target language vocabulary, in both quantitative and qualitative terms. It was necessary to identify the technical register, assess collocational and compositional patterns, and identify hierarchical and equivalency relationships between vocabulary items (see, for example, (Bédard, 1986)). As described in the previous section, we obtained a corpus of some 280 previously translated documents from the customer. This corpus, as well as serving the purposes of mapping rule construction, also served as a raw corpus for vocabulary analysis. Though it didn't ensure complete coverage of the domain, it was highly representative of the text types to be translated by the system.

A certain amount of normalization was possible in the target text, as a result of the corpus analysis undertaken as described above. The process was largely analogous to that done for the source text, and hence need not be elaborated here.

### *6.2.2. Functional recoding*

Target lexical terms are not by themselves useful in the KANT generation context; they must be converted into a data structure which shows the structural properties of these terms. For example, when the term *réglage du palier de fusée de direction* ("steering knuckle bearing adjustment") needs to be combined with the modifier *initial* ("preload"), structural information is required to decide the latter's placement within the host term. Similarly, all inflectable items (e.g., the head and its modifiers) must be structurally identifiable, to allow for inflections like gender or number agreement. (The problem is even more pervasive in morphologically case-marked languages.) Some level of structural information is required to pluralize *clapet de retenue unidirectionnel* ("one-way check valve") as *clapet<u>s</u> de retenue unidirectionn<u>aux</u>* .

The structural description used in the KANT generation system is the f-structure already described above, though for generation purposes, the fs reflects lexical and syntactic properties of the target language, French. Obviously, the target term feature structures vary in complexity with the term itself; many are so complex that hand-encoding becomes prohibitively difficult. For this reason, a more automated approach was used to encode the 65,000 or so French target terms.

To accomplish this term recoding, a complete French parser was developed. It supports the parsing of nominal complexes, including prepositional and phrasal complements. Obviously, the development of the French term parser, with its associated tagged single-word lexicon, was a considerable undertaking. It follows that, in order to parse the French terms in order to express their compositionality, the well-known problems of attachment and scope become a factor. For this purpose an interactive tool, based on a slightly modified version of the generator, was developed to help visualize the structure of candidate f-structures. Consider the following traces for the translations of "lp gas gauge" and "Cat tpms downloading software", respectively:

```
"indicateur de niveau de gaz de pétrole liquéfié"
1. ( ( indicateur liquéfié ) de ( niveau de ( gaz de pétrole ) ) )
2. ( indicateur de ( niveau de ( ( gaz liquéfié ) de pétrole ) ) )
3. ( indicateur de ( niveau de ( gaz de ( pétrole liquéfié ) ) ) )
4. ( indicateur de ( ( niveau liquéfié ) de ( gaz de pétrole ) ) )
Choose? (Y or N): y
"Number? "3
```

```
"logiciel de chargement de l'indicateur de charge utile Cat"
0. ((logiciel cat) de (chargement de ((l'indicateur utile) de charge)))
1. ((logiciel cat) de (chargement de (l'indicateur de (charge utile))))
2. ((logiciel cat) de ((chargement utile) de (l'indicateur de charge)))
3. ((logiciel (utile cat)) de (chargement de (l'indicateur de charge)))
4. (logiciel de (chargement de ((l'indicateur cat) de (charge utile))))
5. (logiciel de (chargement de ((l'indicateur (utile cat)) de charge)))
6. (logiciel de (chargement de (l'indicateur de (charge (utile cat)))))
7. (logiciel de ((chargement (utile cat)) de (l'indicateur de charge)))
8. (logiciel de ((chargement cat) de ((l'indicateur utile) de charge)))
9. (logiciel de ((chargement cat) de (l'indicateur de (charge utile))))
Choose? (Y or N): y
"Number? "1
```

In many cases, a human with some knowledge of the domain is able to choose the correct decomposition, though expert domain knowledge is often required. To attenuate this problem and reduce reliance on expert domain knowledge, another use of the source corpus was devised. Following observations that compositionality patterns are often cross-linguistic in nature, especially between closely-related languages like English and French (see e.g., (Bauer, 1978)), we succeeded in leveraging the source language corpus to produce data for assisting in the target-language decomposition task.

By assuming strictly binary compositional formation (Warren, 1978), we analyzed the source nominal compounds by scoring the subterms at each level of composition. These scores, which reflected frequencies of independent occurrence of each subterm, were accumulated and passed on to the top level with each composition. The end result was a set of one or more descriptions, rated by expected decomposition, for each term. For example, the following term decompositions were obtained in this manner:

```
((bypass valve displacement)
  ((((bypass valve) displacement) 98%)
   ((bypass (valve displacement)) 2%)))

((spray system water tank level)
  (((((spray system) (water tank)) level) 80%)
    ((((spray system) water) (tank level)) 12%)
    (((spray (system water)) (tank level)) 4%)
    ((spray ((system water) (tank level))) 4%)))
```

Often such decompositions were helpful in providing insight to target term decomposition choices, a process in which we felt the human expert should have the power of decision; accordingly, we did not allow the system to select the decomposition automatically. Obviously, other tools such as the monolingual and bilingual browsers were also helpful in supplying the context sometimes needed to decide between target lexical structural descriptions.

Once target-language terms are parsed, their decompositions selected, and the resulting fs double-checked for consistency, they are integrated into the target lexicon. These lexicon entries are then loaded at run-time and used by the mapper in building a target-language fs representation for each interlingua.

### 6.3. Other Lexical Information

Other types of target-language specific lexical information are available to the system during target language generation. For example, cross-categorical shifts are frequently required during translation. Nominalizations, for example, are common in English-French translation of the client's documentation. Since the morphology does not handle French derivational morphology, a table of possible nominalizations is made available to the generator.

```
("auto-guider" ((:NOM ("auto-guidage" M))))
("établir" ((:NOM ("établissement" M))))
("générer" ((:NOM ("génération" F))))
("préférer" ((:NOM ("préférence" F))))
```

Another type of lexically-dependent processing done at generation time is the selection of complementizers for non-finite complements. In many instances, this selection is lexically conditioned, based on the lexical head. Thus, for example, we have *Il m'aide à travailler.* but *Il choisit de travailler.* Rather than hard-code translation of non-finite-clause-introducing complementizers, a template is inserted at the correct place during target mapping, and the generation component performs lexical selection based on the lexical content at the head of the complement.

Other types of lexically-related information, such as verb-aspect compatibility considerations, causative constructions, and reflexivization properties, are present in the system lexicons but also involve structural (i.e., non-lexical) mapping, and aren't discussed here. In conclusion, then, we have seen that the target lexicon is a significant repository of information required to ensure correct generation of target language word and phrase structures, the building blocks for complete sentential fs construction.

### 6.4. Discussion

As the foregoing discussion has illustrated, we sought a thorough, fine-grained representation of the structural properties of the target lexical items. We therefore had to contend with the thorny issues of attachment decisions and representational complexity. Although automated solutions could be brought to bear, the task is still laborious. In searching for possible simplifications, one might seek a less detailed specification of the target terms' syntax. How fully should a term's structure be articulated? Is a full parse really necessary? Perhaps just the identification of crucial words in each term (ones, for example, that may take inflection) and critical insertion sites (for modifiers) might be adequate for text generation purposes. These are questions with interesting implications for the KANT approach.

There is a considerable amount of overlap in the fs representations of the target terms. For example, words like *huile* occur hundreds of times across the target vocabulary, and its full fs is likewise incorporated *in toto* into the fs of each term it participates in. Clearly a larger degree of inter-structure sharing of the f-structures of common words and subphrases would reduce the size of the lexicon.

We have shown how target-term compositionality judgments were facilitated by a corpus of source-term decomposition analyses. Though tools aided the lexicon builder in process, it was still a largely human-intensive one. It should be possible to write a structure-matching tool which would compare subconstituents and their translations, and relate them to proposed target-term composition patterns. This has not been undertaken for the KANT project, mainly because the available on-line target corpus is anticipated to be smaller for the other target languages planned or in progress.

## 7. 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 other domain-oriented corpora and for other languages.

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. The upper bound of the machine-constructed portion of the system was set mostly by the availability of on-line resources including source-language and target-language corpora and lexicons, and by the overlap of these resources with the chosen domain. We have shown throughout that, although specialized computational tools could be brought to bear in handling the data, considerable hand-crafting was also necessary (an unavoidable circumstance in our estimation, even in the long term). As corpus mining techniques improve and as document production practices follow the current trend toward more imposition of structure (via hierarchical and metatextual markup schemes such as SGML), computational techniques should be even more applicable in the future. Similarly, as those who have acquired expertise in a given domain become more accustomed to perceiving their knowledge in terms of hierarchical or even class-distributed relationships, or in ways that can be declaratively specified, they will become more able to contribute directly in the knowledge acquisition 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.).

## Acknowledgments

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 Lonsdale, Jeanne Mier, and Will Walker. This work has also benefitted greatly from our collaboration with Claude Doré of Taurus Translations. We also express our gratitude to our colleagues at Carnegie Group and Caterpillar for their participation in the project.

## Notes

1. For example, extraction of English/French term pairs from translated texts could only account for 12% of the initial CATALYST French lexicon development, because the small number of aligned texts that were available did not cover the entire domain (Mitamura, Nyberg, and Carbonell, 1993).
2. Concept names are prefixed by the gross semantic category class specifiers *O-, *A-, *P-, *M-, and *U- to represent objects, actions, properties, manners, and units respectively.
3. For the sake of expedience, simple heuristics were utilized for constructing likely noun phrases based on the part of speech assignments in the Brown Corpus. It is possible that more accurate results could be achieved through the use of more involved methods, such as those presented in (Church, 1988; Bourigault, 1992; Chen and Chen, 1994).
4. A tagged version of the Brown Corpus was utilized as a resource for part-of-speech tags.
5. This approach can be contrasted with the approach taken in (Grishman, Macleod, and Meyers, 1994), which makes use of on-line resources to enhance lexicons which are created manually.

## References

Bauer, L., editor. 1978. *The Grammar of Nominal Compounding with Special Reference to Danish, English, and French.* Odense University Press.

Bédard, C. 1986. La Traduction Technique: Principes et Pratique. *Linguatech.*

Bourigault, D. 1992. Surface Grammatical Analysis for the Extraction of Terminological Noun Phrases. In *Proceedings of COLING-92.*

Brown, R.D. 1991. Automatic and Interactive Augmentation. In *A Case Study in Knowledge-Based Machine Translation.* Morgan Kaufmann, San Mateo, CA.

Chen, K. and H. Chen. 1994. Extracting Noun Phrases form Large-Scale Texts: A Hybrid Approach and Its Automatic Evaluation. In *Proceedings of ACL-94*.

Church, K. 1988. A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text. In *Proceedings of the Second Conference on Applied Natural Language Processing*.

Francis, W. and H. Kučera, editors. 1982. *Frequency Analysis of English Usage*. Houghton Mifflin, Boston, MA.

Galinski, C. 1988. Advanced Terminology Banks Supporting Knowledge-Based MT. In D. Maxwell, K. Schubert, and A. Witkam, editors, *New Directions in Machine Translation*. Foris.

Goodman, K. and S. Nirenburg, editors. 1991. *A Case Study in Knowledge-Based Machine Translation*. Morgan Kaufmann, San Mateo, CA.

Grishman, R., C. Macleod, and A. Meyers. 1994. Comlex Syntax: Building a Computational Lexicon. In *Proceedings of COLING-94*.

Leavitt, J., D. Lonsdale, K. Keck, and E. Nyberg. 1994. Tooling the Lexicon Acquisition Process for Large-Scale KBMT. In *Proceedings of IEEE Tools for AI*.

Lonsdale, D. 1994. Extraction d'un Vocabulaire Bilingue: Outils et Méthodes. In A. Clas and P. Bouillon, editors, *Actes du Colloque Lexicologie, Terminologie et Traduction*. Les Presses de l'Université de Montréal.

Lonsdale, D., A. Franz, and J. Leavitt. 1994. Large-scale Machine Translation: An Interlingua Approach. In *Proceedings of IEA/AIE-94*.

Mitamura, T. 1989. *The Hierarchical Organization of Predicate Frames for Interpretive Mapping in Natural Language Processing*. Ph.D. thesis, University of Pittsburgh.

Mitamura, T. and E. Nyberg. 1992. Hierarchical Lexical Structure and Interpretive Mapping in MT. In *Proceedings of COLING 1992*, Nantes, France, July.

Mitamura, T., E. Nyberg, and J. Carbonell. 1991. An Efficient Interlingua Translation System for Multi-lingual Document Production. In *Proceedings of Machine Translation Summit III*, Washington, DC, July.

Mitamura, T., E. Nyberg, and J. Carbonell. 1993. Automated Corpus Analysis and the Acquisition of Large, Multi-Lingual Knowledge Bases for MT. In *Proceedings of TMI-93*.

Nyberg, E. and T. Mitamura. 1992. The KANT System: Fast, Accurate, High-Quality Translation in Practical Domains. In *Proceedings of COLING-92*, Nantes, France, July.

Nyberg, E., T. Mitamura, and J. Carbonell. 1994. Evaluation Metrics for Knowledge-Based Machine Translation. In *Proceedings of COLING-94*.

Tomita, M., M. Kee, T. Mitamura, and J. Carbonell. 1987. Linguistic and Domain Knowledge Sources for the Universal Parser Architecture. In H. Czap and C. Galinski, editors, *Terminology and Knowledge Engineering*. INDEKS Verlag, Frankfurt, Germany, pages 191–203.

Tsujii, T. 1988. What is a Cross-linguistically Valid Interpretation of Discourse? In D. Maxwell, K. Schubert, and A. Witkam, editors, *New Directions in Machine Translation*. Foris.

Warren, B. 1978. Semantic Patterns of Noun-Noun Compounds. Technical report, Acta Universitatis Gothoburgensis.