[From: *Georgetown University Round Table on Languages and Linguistics 1989*, James E.Alatis, ed.]

New developments in knowledge-based machine translation

Sergei Nirenburg
Carnegie Mellon University

Recent experience in Knowledge-Based Machine Translation (KBMT) suggests that systems adhering to this paradigm can be useful not only in very small domains and limited sublanguages. With special attention paid to acquisition of large knowledge bases and with the advent of new tools (including representation languages, human-computer interfaces, database management systems, etc.) the practicality of the knowledge-based approach is steadily growing. Since a comprehensive automatic analysis of meaning is not yet feasible, and the attainment of this goal will remain the central objective of computational linguistics for years to come, a practical KBMT system will be necessarily of a hybrid nature. It may include certain types of processing that would be considered more appropriate for a transfer-oriented system. It will also include a measure of human involvement. However, while in transfer systems human involvement invariably means postediting, human-aided knowledge-based systems will use human help *during* the process of translation, to finalize the representations of the meanings in a source language (SLG) text. It is expected that the target language (TL) texts produced from such improved meaning representations will be of comparable quality with translations produced by humans.

1. 'Transfer or interlingua?'--Is this question still relevant? Historically, machine translation (MT) systems have been of three major types: direct, transfer, and interlingua. Detailed descriptions of the three approaches, with all their modifications and varieties, can be found in the MT literature (see, in particular, Hutchins 1986, Zarechnak 1979). Direct systems have been justly criticized for their *ad hoc*ness, so that at present the choice of architectures for machine translation systems is reduced to the two latter approaches. In this section we will very briefly comment on the essential differences between them and suggest that the latter are possibly less important methodologically than the altitude to meaning analysis and also coverage.

Transfer systems involve a measure of target language-independent analysis of the source language. This analysis is usually syntactic. It allows substituting SLG lexical units with TL lexical units *in context.* That is, it permits taking into account the types of syntactic sentence constituents in which lexical units appear.

In interlingua systems the SLG and the TL, are never in direct contact. The processing in such systems has traditionally been understood to involve two major stages: representing the meaning of a SLG text in an artificial formal language, *interlingua,* and then expressing this meaning using the lexical units and syntactic constructions of the target language. Few interlingua systems have been fully implemented because of the very significant complexity (both theoretical and empirical) of extracting a 'deep' meaning from a natural language text.

The major distinction between the interlingua- and the transfer-based systems is, in fact, not so much the presence or absence of a bilingual lexicon but rather the attitude towards comprehensive analysis of meaning. In practice, those MT researchers who believe in translating without 'deep' understanding (or perhaps who believe in the unattainability of 'deep' understanding) of the SLG text tend to prefer the transfer paradigm. The price they have to pay for avoiding meaning analysis is the need for an extra step in the translation process, namely, postediting.

Inherently, a transfer system can involve many levels of meaning analysis. This becomes clear when one considers that different transfer-based systems have widely varying levels at which transfer occurs—from simple phrase structure trees to detailed representations that use subcategorization patterns, and even selectional restrictions. There is a trend in transfer-based MT to downplay the necessity of structural transfer, that is, the stage of transforming standard syntactic structures of the SLG into the corresponding TL structures. This is in part due to the prevalence of grammatical theories that eschew transformations and seek universally applicable representations of grammatical structures and relations. This trend is essentially interlingual in nature. Transfer-based systems can also deal with lexical semantics; the language in which the meanings of SLG lexical units are expressed will be the TL itself. This can be implemented through a bilingual lexicon featuring disambiguation information.

In interlingua systems the meanings are represented in an artificial language—the reason being that such a language is better suited for the formulation of disambiguation rules necessary for producing an adequate meaning of a SLG text, in part because it was specifically designed for this purpose.[1]

Distinctions between the transfer and the interlingual approaches are best drawn at an abstract level. In reality, when practical systems have to be built, many types of work will be practically identical for both approaches (notably, the grammars and programs for syntactic analysis and synthesis). For some other types of work the very nature of the material dictates the necessity of methodological compromises—for instance, some source language lexical units for which the interlingua does not, at the moment, have an adequate representation can be treated in a transfer-like manner in a practical knowledge-based machine translation (KBMT) system. At the same time, for those (very frequent) cases when there is no possibility of direct transfer of a lexical unit or a syntactic structure between two languages, a transfer system would benefit by trying to express the meaning of such lexical units and syntactic structures and construct a TL correlate from this (more detailed and transparent) representation. The requirements of practical use, indeed, pose similar difficulties for both approaches—consider such universal problems as ill-formed input, special symbols and codes, document layout preservation, translatable material in figures, etc. ATLAS-II (Uchida 1987) is an example of a hybrid MT system that has features from both major approaches.

2. New features in KBMT. I will illustrate the recent progress in knowledge-based machine translation using as the example KBMT-89 (Nirenburg and Goodman 1989), a system recently developed at the Center for Machine Translation of CMU. It translates from English into Japanese and from Japanese into English in the domain of computer hardware installation manuals. Small-scale extensions are being developed at present to add French, Russian, and Polish to the list of source languages. The system development was sponsored by IBM.

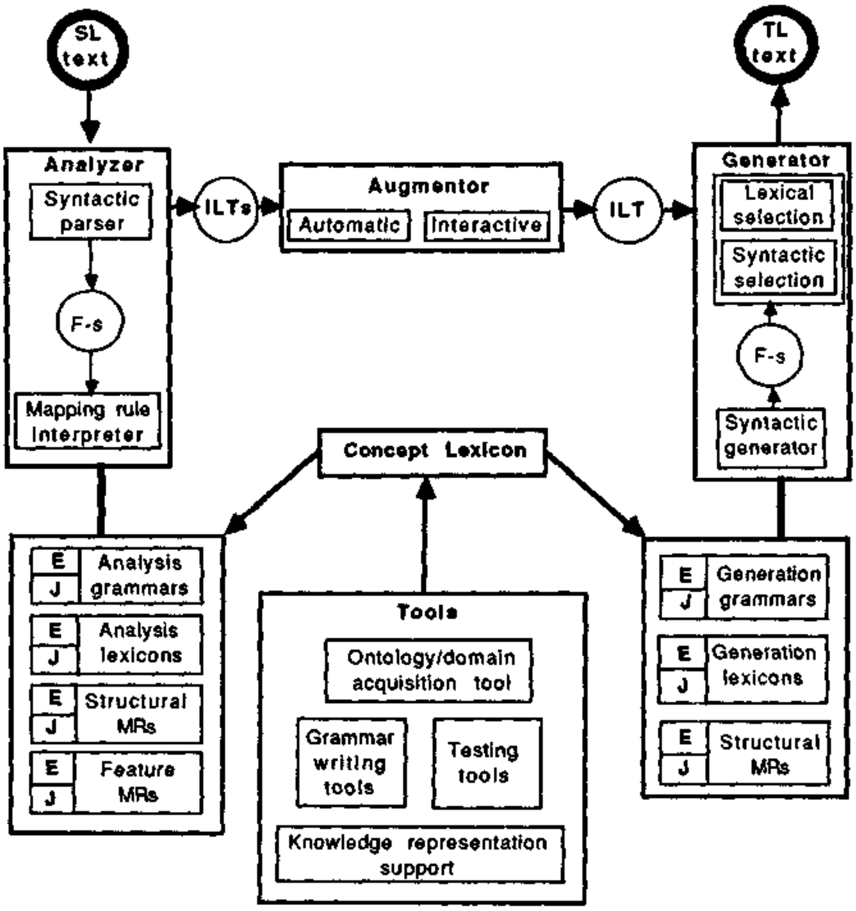
KBMT-89 consists of a large number of modules, as illustrated in Figure 1. There are four grammars; two analysis lexicons and two generation lexicons; a domain model (called 'concept lexicon'); an augmentor, which serves as a filter between analysis and generation and a set of acquisition, maintenance, and testing tools. In terms of coverage and inventory of computational and knowledge components, KBMT-89 appears to be the largest KBMT system to date. For a sketch of the global architecture of the system, see Figure 1: The global architecture of KBMT-89.

In what follows I will briefly discuss several of the distinguishing features of KBMT-89. The system has many more salient features. See Nirenburg and Goodman 1989, for a more detailed discussion.

2.1 Nature and size of knowledge bases. KBMT-89 is supported by a domain model of about one thousand concepts organized into a complex network. As can be seen from Figure 2, the knowledge base used in KBMT-89 is capable of representing not only general domain knowledge about taxonomies of object types (e.g., 'a car is a kind of a vehicle,' 'a doorhandle is a part of a door,' 'artifacts are characterized (among other properties) by the property *made-by*') but also knowledge about particular instances of object types (e.g. IBM can be included into the domain model as a marked instance of the object type 'corporation') as well as instances of (potentially, complex) event types[2] (e.g. the election of George Bush as President of the United States is a marked instance of the complex action 'to-elect'). The ontological part of the knowledge base takes the form of a multihierarchy of concepts connected through taxonomy-building links, such as *is-a*, *part-of* and some others. We call the resulting structure a multihierarchy because concepts are allowed to have multiple parents on a single link type.

In KBMT-89 the ontological concepts are first subdivided into objects, events, forces (introduced to account for intentionless agents) and properties. Properties are further subdivided into relations and attributes. Relations are mappings among concepts (e.g. *belongs-to* is a relation, since it maps an object into the set {*human *organization}) while attributes are mappings of concepts into specially defined value sets (e.g. *temperature* is an attribute that maps physical objects into values on the semiopen scale '> *0',* with the granularity of one degree Kelvin). Concepts are represented as frames whose slots are properties fully defined in the system.

Figure 1. The global architecture of KBMT-89.



In order to acquire and maintain the KBMT-89 domain model we developed an interactive knowledge acquisition and maintenance system, ONTOS (for a detailed discussion see e.g. Nirenburg et al. 1988a). To illustrate the way ONTOS operates, consider Figures 3 and 4. They show the ONTOS screen at two strategic moments during the acquisition of the concept of the Macintosh microcomputer. The acquisition is performed by using and extending the information already available in the domain model about the parents and (if available) siblings of the new concept.

Many models of a single domain are possible, and it is methodologically improper to insist on there existing a single canonical domain model. The set of ontological postulates used in KBMT-89 (and illustrated here) has been deliberately made as general as possible in order to make it adaptable to other views of the world. Using the KBMT-89 acquisition tools and, optionally, using the KBMT-89 domain model as the starting point, other researchers can build their own domain models within a short period of time—the task certainly impossible before the advent of knowledge acquisition and maintenance systems.

Figure 2. Metatypes of entities and relations in KBMT-89 domain model. The domain model also serves as an index into the analysis and generation lexicons for both English and Japanese. It represents both ontology (semantic memory) and experience (episodic memory).

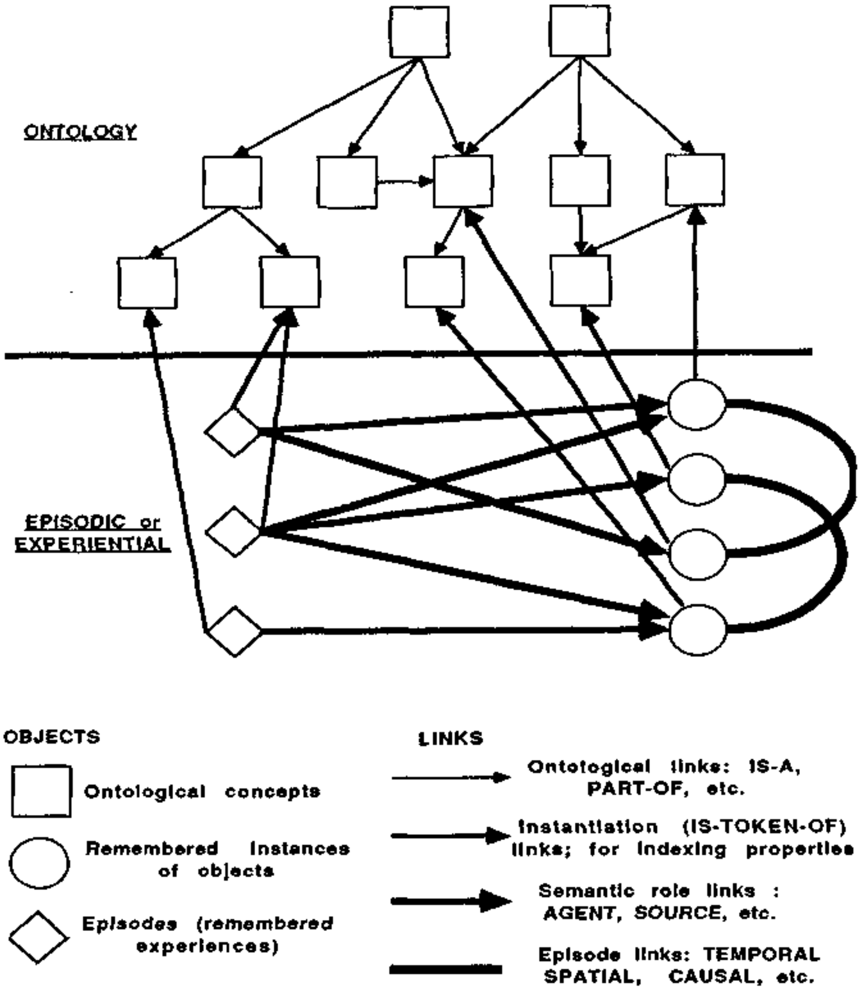
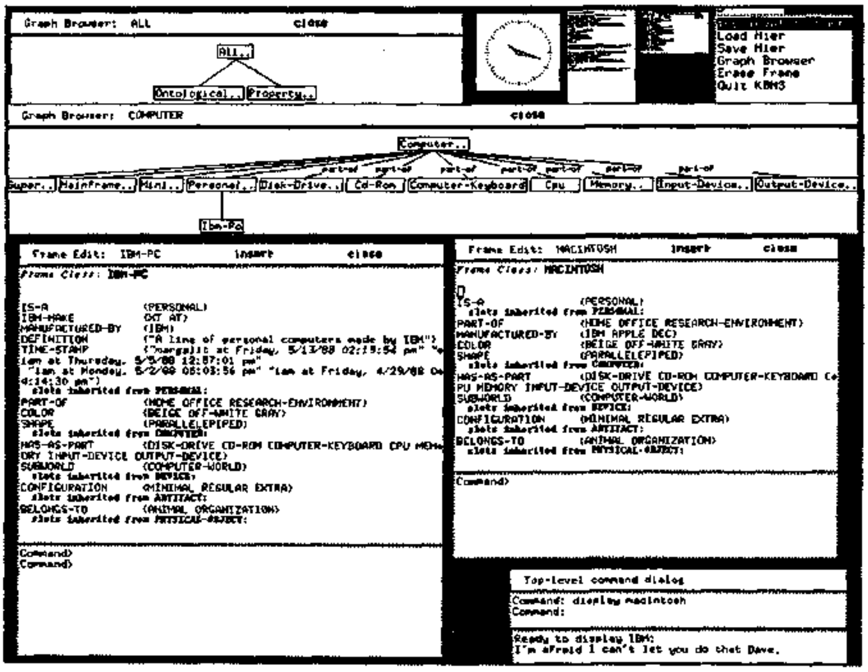

ONTOLOGY
EPISODIC or EXPERIENTIAL
OBJECTS
Ontological concepts
Remembered instances of objects
Episodes (remembered experiences)
LINKS
Ontological links: IS-A, PART-OF, etc.
Instantiation (IS-TOKEN-OF) links; for indexing properties
Semantic role links : AGENT, SOURCE, etc.
Episode links: TEMPORAL SPATIAL, CAUSAL, etc.

Figure 3. The concept 'macintosh computer' in its initial form.



```
Frame Class: IBM-PC

IS-A                (PERSONAL)
IBM-MAKE            (XT AT)
MANUFACTURED-BY     (IBM)
DEFINITION          ("A line of personal computers made by IBM")
TIME-STAMP          ("margalit at Friday, 5/13/88 02:15:54 pm" "e
ian at Thursday, 5/5/88 12:57:01 pm"
 "ian at Monday, 5/2/88 05:03:56 pm" "ian at Friday, 4/29/88 0
4:14:30 pm")
  slots inherited from PERSONAL:
PART-OF             (HOME OFFICE RESEARCH-ENVIRONMENT)
COLOR               (BEIGE OFF-WHITE GRAY)
SHAPE               (PARALLELEPIPED)
  slots inherited from COMPUTER:
HAS-AS-PART         (DISK-DRIVE CD-ROM COMPUTER-KEYBOARD CPU MEM
ORY INPUT-DEVICE OUTPUT-DEVICE)
SUBWORLD            (COMPUTER-WORLD)
  slots inherited from DEVICE:
CONFIGURATION       (MINIMAL REGULAR EXTRA)
  slots inherited from ARTIFACT:
BELONGS-TO          (ANIMAL ORGANIZATION)
  slots inherited from PHYSICAL-OBJECT:

Command>
Command>
```

```
Frame Class: MACINTOSH

IS-A                (PERSONAL)
  slots inherited from PERSONAL:
PART-OF             (HOME OFFICE RESEARCH-ENVIRONMENT)
MANUFACTURED-BY     (IBM APPLE DEC)
COLOR               (BEIGE OFF-WHITE GRAY)
SHAPE               (PARALLELEPIPED)
  slots inherited from COMPUTER:
HAS-AS-PART         (DISK-DRIVE CD-ROM COMPUTER-KEYBOARD C
PU MEMORY INPUT-DEVICE OUTPUT-DEVICE)
SUBWORLD            (COMPUTER-WORLD)
  slots inherited from DEVICE:
CONFIGURATION       (MINIMAL REGULAR EXTRA)
  slots inherited from ARTIFACT:
BELONGS-TO          (ANIMAL ORGANIZATION)
  slots inherited from PHYSICAL-OBJECT:

Command>
```

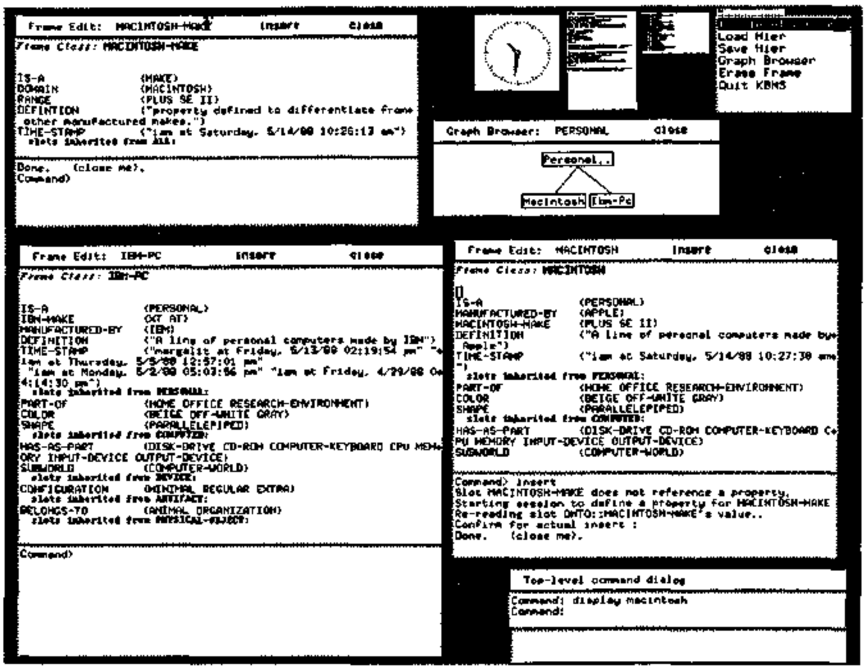
```
Command: display macintosh
Command:

Ready to display IBM:
I'm afraid I can't let you do that Dave.
```

2.2 **Exploring additional types of meaning.** Traditionally, the types of meanings represented in knowledge-based natural language processing systems have been almost entirely propositional. A large number of nonpropositional—pragmatic and discourse-meanings, such as thematic structure, speech act, modality, discourse cohesion, speaker attitude, etc. were not overtly represented. In knowledge-based machine translation such meanings have been traditionally ignored. In transfer systems they are treated implicitly, based on the observation that often no special processing has to be done other than simple transfer of lexical clues or, sometimes, a structural transformation. This approach is prone to error and does not support any paraphrasing capability necessary when text-level stylistic decisions are treated.

We have argued elsewhere (Nirenburg and Carbonell 1987) for the necessity of incorporating nonpropositional knowledge into the representation of the results of text analysis (known as interlingua text in KBMT-89). Figure 5 illustrates how both propositional (domain model-dependent) and nonpropositional knowledge are integrated in a single representation. Analysis lexicon entries are correspondingly classified into those mapping into instances of domain model concepts and those signifying nonpropositional properties and thus mapping directly into specific property values in ILT.

Figure 4. Adding properties on the basis of inheritance and sibling differentiation.



```
Frame Edit:  MACINTOSH-MAKE          insert          close
Frame Class: MACINTOSH-MAKE

IS-A                  (MAKE)
DOMAIN                (MACINTOSH)
RANGE                 (PLUS SE II)
DEFINTION             ("property defined to differentiate from
 other manufactured makes.")
TIME-STAMP            ("iam at Saturday, 5/14/88 10:26:13 am")
  slots inherited from ALL:

Done.    (close me).
Command>
```

```
Frame Edit:  IBM-PC          insert          close
Frame Class: IBM-PC

IS-A                  (PERSONAL)
IBM-MAKE              (XT AT)
MANUFACTURED-BY       (IBM)
DEFINITION            ("A line of personal computers made by IBM")
TIME-STAMP            ("margalit at Friday, 5/13/88 02:19:54 pm" "
iam at Thursday, 5/5/88 12:57:01 pm"
 "iam at Monday, 5/2/88 05:03:56 pm" "iam at Friday, 4/29/88 0
4:14:30 pm")
  slots inherited from PERSONAL:
PART-OF               (HOME OFFICE RESEARCH-ENVIRONMENT)
COLOR                 (BEIGE OFF-WHITE GRAY)
SHAPE                 (PARALLELEPIPED)
  slots inherited from COMPUTER:
HAS-AS-PART           (DISK-DRIVE CD-ROM COMPUTER-KEYBOARD CPU MEM
ORY INPUT-DEVICE OUTPUT-DEVICE)
SUBWORLD              (COMPUTER-WORLD)
  slots inherited from DEVICE:
CONFIGURATION         (MINIMAL REGULAR EXTRA)
  slots inherited from ARTIFACT:
BELONGS-TO            (ANIMAL ORGANIZATION)
  slots inherited from PHYSICAL-OBJECT:

Command>
```

```
Frame Edit:  MACINTOSH          insert          close
Frame Class: MACINTOSH

IS-A                  (PERSONAL)
MANUFACTURED-BY       (APPLE)
MACINTOSH-MAKE        (PLUS SE II)
DEFINITION            ("A line of personal computers made by
 Apple")
TIME-STAMP            ("iam at Saturday, 5/14/88 10:27:38 am
")
  slots inherited from PERSONAL:
PART-OF               (HOME OFFICE RESEARCH-ENVIRONMENT)
COLOR                 (BEIGE OFF-WHITE GRAY)
SHAPE                 (PARALLELEPIPED)
  slots inherited from COMPUTER:
HAS-AS-PART           (DISK-DRIVE CD-ROM COMPUTER-KEYBOARD C
PU MEMORY INPUT-DEVICE OUTPUT-DEVICE)
SUBWORLD              (COMPUTER-WORLD)

Command> insert
Slot MACINTOSH-MAKE does not reference a property.
Starting session to define a property for MACINTOSH-MAKE
Re-reading slot ONTO::MACINTOSH-MAKE's value..
Confirm for actual insert :
Done.    (close me).
```

```
Top-level command dialog
Command: display macintosh
Command:
```

2.3 Focus on generation. Unlike most machine translation systems, KBMT-89 pays a significant amount of attention to the generation side of the process. To give just one example, let me illustrate one component of the generation process—lexical selection. When the knowledge-based approach is used, it becomes possible to enhance the process of lexical selection (lexical synonymy resolution) in generation. Figure 6 shows lexical selection steps used in KBMT-89.[3] Note that the lexical selection process involves filters that are essentially syntactic and source-language dependent (such as subcategorization) as well as semantic filters (such as the meaning matching metric, which operates on language-independent meaning representations) and stylistic fillers (for instance, the preference, while generating English, for a verb to realize the meaning of the head of a proposition).

2.4 Human-computer interaction. The idea of human-aided machine translation occurred to MT researchers very early. Of a number of ways in which humans can facilitate the process of automatic translation we are mostly interested in having a human user verify, improve, and finalize the system's decisions during analysis. The system may be unable (have no knowledge) to prefer one candidate reading of the input over another. Or, alternatively, its knowledge may rule out all of the candidate readings. Human intervention may become necessary.

Figure 5. The interaction among lexicons and ILT. Note that some source language lexical units are connected to their interlingua meanings directly, bypassing the Concept Lexicon. The figure also illustrates the lack of symmetry in the treatment of lexical semantics in analysis and generation; the main problem in analysis is polysemy, while in generation it is synonymy.

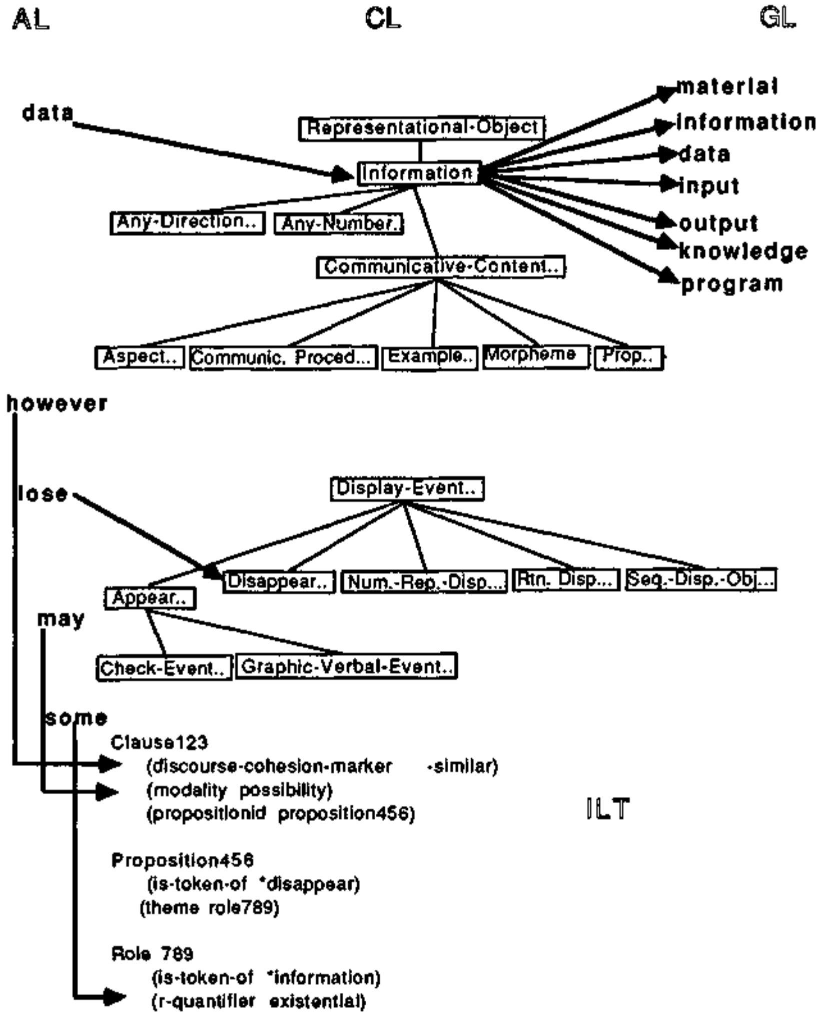
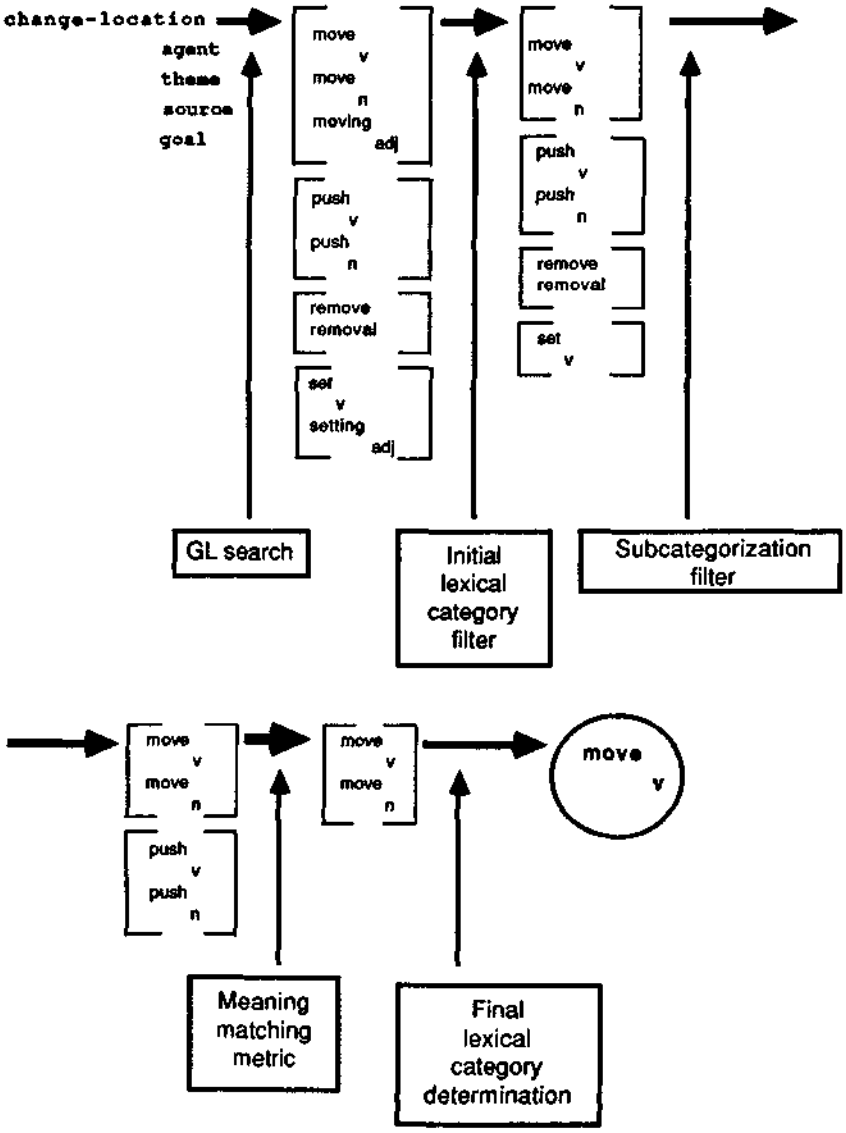


Figure 6. An illustration of the process of lexical selection.

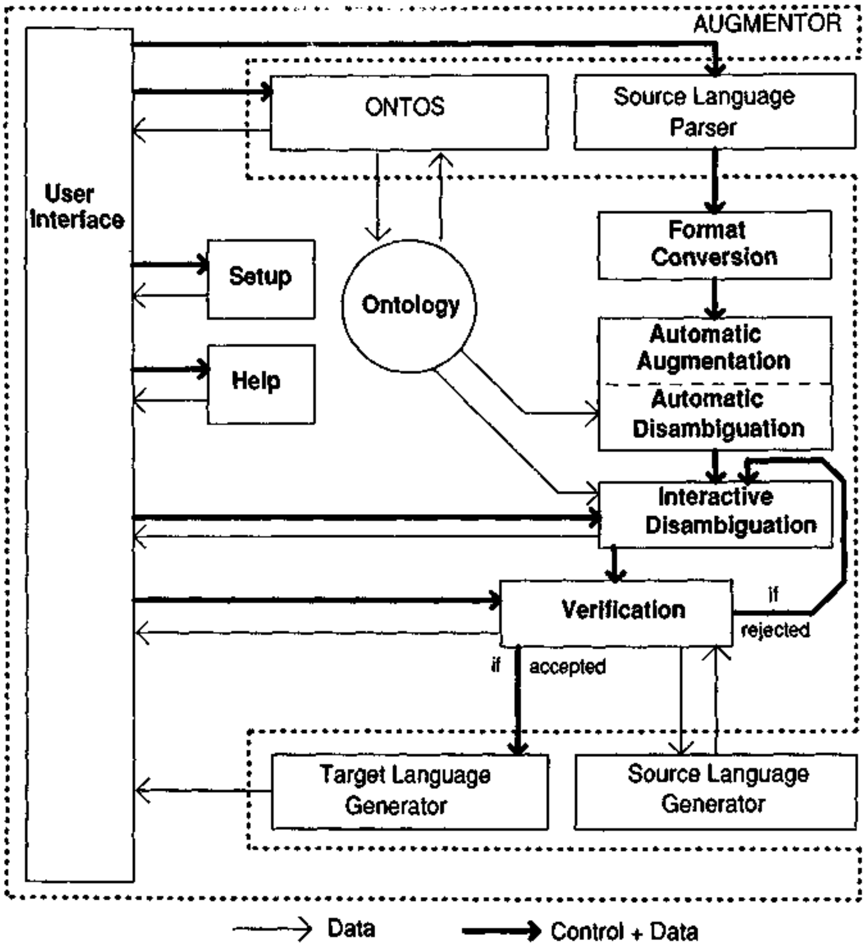


The human-computer interface that would support the interaction of the system and the human user, the interactive editor, has been implemented on a small scale in the MIND machine translation project at RAND Corporation (Kay 1973; the module was called 'disambiguator'). In KBMT-89 the interactive module (known as the 'augmentor') has been implemented on a larger scale.

In fact, the augmentor serves not only as an interface. It is a general filter between analysis and generation. It has an automatic component, which in KBMT-89 deals with referential ambiguity resolution and assignment of non-propositional meanings. The augmentor can also be used for knowledge format modifications that may become necessary if an independently developed analysis module is integrated with the system. The interactive component of the augmentor queries the user about the residual lexical ambiguities, residual problems in attachment of prepositional phrases and subordinate clauses, properties on which nominal modifiers are linked to the heads in noun-noun compounds, etc.

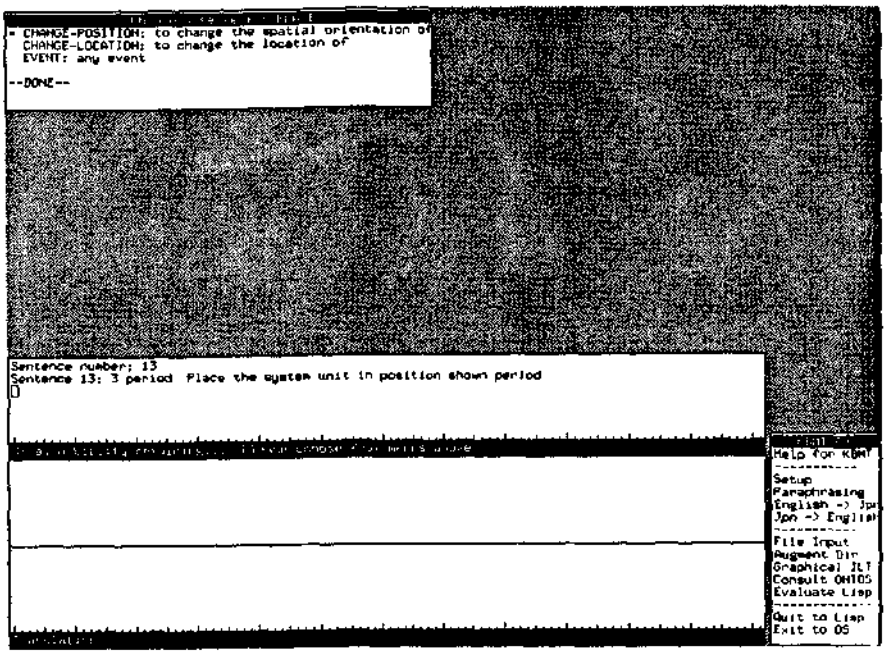
Figure 7 shows the architecture of the KBMT-89 augmentor, while Figure 8 presents a sample augmentor session.

Figure 7. A representation of the Augmentor architecture.



Since the augmentor is integrated in the general user interface of the entire system, Figure 8 also shows the view of the screen during system operation.

Figure 8. Choosing among the remaining candidates.



```
■ CHANGE-POSITION: to change the spatial orientation of
  CHANGE-LOCATION: to change the location of
  EVENT: any event

--DONE--
```

```
Sentence number: 13
Sentence 13: 3 period  Place the system unit in position shown period
```

2.5 MT as an experimental testbed for computational linguistics. In addition to its utility as a machine translation shell, KBMT-89 (or a similar KBMT system) can be very profitably used as a research tool and testbed in computational linguistics and artificial intelligence. To illustrate briefly:

- In its current state the system provides an excellent tool for devising and testing new and more powerful specialized semantic interpretation algorithms, such as, for instance, noun-noun compound under standing or prepositional phrase attachment. With more types of semantic and pragmatic knowledge appearing in the ILT, more specialized 'microtheories'[4] will be devised and/or incorporated into the process.

- The generation component of KBMT-89 is a very good substrate on which to build more sophisticated natural language generators. In particular, it facilitates the interaction of syntactic, lexical and prosodic processing and offers a level of reliance on world knowledge that is unusual in most current natural language generators.

- An additional advantage of using KBMT systems as research vehicles is that it is a *comprehensive* system that allows immediate testing of a new component, such as a new parser or a generator, in the context where a 'real' output can be obtained.

- The interface component of a KBMT can serve as a medium for building other interfaces, notably for the purpose of computer-aided instruction and, in particular, for teaching foreign languages. The interface can also be very useful in machine learning systems, especially those studying learning from text or learning by being told, or in systems that investigate hybrid learning processes which involve natural language.

- A comprehensive understanding-and-generation system like KBMT-89 can also be used as a component in a system modelling a cognitive agent—alongside other modules, such as planning and problem solving, perception and action simulation components.

- The ontological and domain knowledge in KBMT-89 can serve as a tool for research in the area of acquisition and maintenance of large knowledge bases. In fact, Ontos is already being used to build domain models in the fields of molecular biology, law, financial transactions and computer software in the framework of projects in the areas of diagnostic expert systems, qualitative process theory and computer-aided instruction. The domain models can also serve as the underlying substrate for a hypertext-type index into a large corpus of human-readable information.

- The computing technology embodied by a KBMT system can be used in other applications. One of the areas in which KBMT-89 can yield immediate practical results is design and development of high-quality translator's workstations. The interaction environment can be extended to include additional types of human-computer interaction. Additional knowledge sources can be connected to the system (for instance, human-readable dictionaries and encyclopedias). And the presence of working analyzer and generator modules will allow the system to suggest acceptable solutions (or informed choices) to the human translator; this is a feature not present in any current translator's workstation.

- Outside of machine translation proper, the technology developed in KBMT-89 is readily usable in applications that require different types of inputs and/or outputs to a natural language processor. Thus, instead of forwarding an intralingua text to the generator, one can pass it on to a special reasoning program that will produce an abstract of the input text, or answer questions based on it, or categorize the input text into one of a number of taxonomic classes. KBMT-89 (and other knowledge-based machine translation systems) can also be reconfigured for supporting natural language interfaces to database systems. Indeed, if a data manipulation (query) language is substituted for the interlingua, the task of query formulation can become quite similar to that of analyzing a natural language input for translation.

Notes

[1] A very good example of what happens when an interlingual system chooses a human-oriented language as the interlingua is the DLT project (Witkam 1987). This project has selected Esperanto as the interlingua and ended up with the necessity of a significant overhaul of the language to make it support the types of processing than an interlingua must support.

[2] Such event instances are sometimes called 'episodes'. Their inclusion, together with object instances, into a domain model is a precondition for designing systems that would automatically augment domain models, based on the experience of reading and understanding texts. Such systems can be included as components in future knowledge-based machine translation systems.

[3] The KBMT-89 generator is actually a subset of DIOGENES-88, a distributed natural language generator developed at CMU (see Nirenburg et al. 1988). The latter system involves even more knowledge in the lexical selection process (in particular, the knowledge of language-dependent lexical collocations is used; see Nirenburg and Nirenburg 1988 for a description).

[4] The concept of microtheory that we use here has been described in greater detail in Nirenburg and Pustejovsky (1988). Provided that a comprehensive, computationally relevant theory of semantic and pragmatic interpretation is not about to be advanced, the best policy for computational linguists in building comprehensive, even though sublanguage-dependent, natural language processing systems is to combine, to the best of their ability, the results offered by partial theories (or, *microtheories)* of particular semantic and pragmatic phenomena (e.g. quantification, reference, thematic structure, discourse cohesion, aspect, time and tense, metaphor, metonymy, etc.). Under this approach, we can use, we can operate without an integrated semantic and pragmatic theory, once we make sure that the findings are represented in a uniform way and introduce a computational control structure that will allow a high degree of autonomy to the component microtheory-based modules, while at the same time maintaining interdependence of microtheory-dependent choices.

References

Hutchins, W. 1986. Machine translation: Past, present, future. Chichester, U.K.: Ellis Horwood Ltd.

Kay, M. 1973. The MIND system. In: R. Rustin, ed., Natural language processing. New York: Algorithmics Press.

Nirenburg, S., and J. Carbonell. 1987. Integrating discourse pragmatics and propositional knowledge in multilingual natural language processing, vol. 2.2. Montreal: The Cognitive Science Society.

Nirenburg, S., I. Monarch, T. Kaufmann, I. Nirenburg. and J. Carbonell. 1988a. Acquisition of very large knowledge bases. Technical report, Center for Machine Translation, Carnegie Mellon University.

Nirenburg, S., R. McCardell, E. Nyberg, P. Werner, S. Huffman, E. Kenschaft, and I. Nirenburg. 1988b. DIOGENES-88. Technical Report, Center for Machine Translation, Carnegie Mellon University.

Uchida, H. 1987. ATLAS: Fujitsu machine translation system. Proceedings of MT Summit, Tokyo, September.

Witkam, T. 1987. Interlingual MT—an industrial initiative. Proceedings of MT Summit, Tokyo, September.

Zarechnak, M. 1979. The history of machine translation. In: B. Henisz-Dostert, R. Ross Macdonald, and M. Zarechnak, eds., Machine translation. Trends in linguistics: Studies and monographs, vol. 11. The Hague: Mouton.