Machine Translation Again?

Yorick Wilks, Jaime Carbonell, David Farwell, Eduard Hovy and Sergei Nirenburg
Department of Computer Science
New Mexico State University
Las Cruces, NM 88001

Machine translation (MT) remains the paradigm task for natural language processing (NLP) since its inception in the 1950s. Unless NLP can succeed with the central task of machine translation, it cannot be considered successful as a field. We maintain that the most profitable approach to MT at the present time is an interlingual and modular one. MT is one the precious few computational tasks falling broadly within artificial intelligence (AI) that combine a fundamental intellectual research challenge with enormous proven need. To establish the latter, one only has to note that in Japan alone the current MT requirement is for 20 billion pages a year (a market of some $66 billion a year).

The vulgarized version of the history of MT is as follows: In the 1950s and 1960s large funds were made available to US MT which proved to be an umitigated failure. The ALPAC report (1966) said MT was impossible and doomed all further US funding. MT work then moved to Canada and Europe where it partly succeed, which was then followed by highly successful exploitation in Japan. The truth, of course, is not at all like that.

MT work did not stop in the US after ALPAC: the AFOSR continued to fund it in the US and there were and are enormous commercial developments (the best known systems being SYSTRAN, ALPS, LOGOS, METAL and SMART).

ALPAC did not say MT was impossible nor that the work done was no good: only that at that point history, with the cost and power of 1960s computers, human translation was arguably cheaper.

MT work did not really move to Europe, since it stopped there also in response to the ALPAC report. The UK believed the ALPAC report, and only in France did serious work continue, and the GETA system in Grenoble became the foundation for a range of others, including the major Japanese university system (Mu) and aspects of the Eurotra system, which was designed to be a multilingual system between the languages of the EEC.

The GETA system, like SYSTRAN, date their origins from the very earliest period of pre-ALPAC MT. The longevity of such systems is proof of the need of stamina and persistence in MT to achieve serious results, but also the need for periodic redesign, pretty much from scratch, since old formalisms and software reach a point where they cannot be further optimized, a point reached long ago with SYSTRAN itself. One way in which all MT work is in SYSTRAN's debt is that it is the main existence proof: it convinces doubters that there that machine translation now exists, albeit in primitive form, and can be purchased on a large scale and at a quality that many users find acceptable for their needs. A key defect in the ALPAC report was that it underestimated how large a market there was for partially accurate, low quality MT, and SYSTRAN filled that market. The point now, of course, is to move on to the huge market for higher-quality MT. But even now the proportion of internal EEC documentation translation for which a preliminary draft version is done by machine is rapidly growing using a modified version of the SYSTRAN system.

It is certainly not the case that most major MT installations in the world are now Japanese. In the list given in the JEIDA report only one Japanese system occurs among the list of major installed systems in the world outside Japan. All the rest are American. However, that list is becoming quickly dated, as Japanese system are being researched, developed and deployed at a much faster rate, reflecting a lopsided ten-to-one total R funding skew in favor of Japan over America. Moreover, some commercial American MT efforts are being purchased by the Japanese; witness BRAVIS's purchase of WEIDNER COMMUNICATIONS, and a (partial) purchase of SYSTRAN. A crucial difference between US and foreign strategies to date has been that the Japanese government made machine translation central to the Fifth Generation effort, and the European Community began ten years ago a $45 million investment in the Eurotra project as part of their overall information technology drive.

Why this is a good time to get back into MT.

There is a growing need for translation in intelligence, commerce, science, government, and international organizations. This is due to factors such as the following:

- Increases in international cooperation and competition, which involve an ever-growing volume of text to be communicated.
- World-wide electronic networks have made international communication much easier.
- Reports, documentation, legal papers, and manuals are increasingly produced in one culture and exported to various other cultures, often in multiple languages.
- More emphasis is being placed on the use of national languages in documents and systems.
- The economic rise of South-East Asia and the opening of the European market in 1992 add significantly to these factors.

Strategic Reasons for an MT Effort:

MT systems live and decay like natural organisms: they have natural life spans that cannot be indefinitely prolonged. The SYSTRAN system has lived long and done well but it is 30 years old and cannot be optimized above the 75% level. Later systems from the early 1970s (GETA, LOGOS, ALPS, WEIDNER, MU, etc) were better constructed but cannot rise above their current levels---the evidence for this being that the two research systems in that list (GETA & MU) have now effectively collapsed and their teams dispersed. The right thing is now to promote a new design using the enormous and transferable advances that have been made in interfaces, hardware, linguistics, AI, machine lexicons etc.

The most recent new large-scale efforts have either been badly managed and proved impractical (like EUROTRA) or set very narrow commercial goals (usually involving only Japanese and English or Asian languages) like the major Japanese systems.

The need has never been greater, not only for MT itself but all the associated technologies that can be integrated into a well-designed MT system. The rapid integration of the world is not leading to a common language (English) nearly as fast as it is leading to the absolute need to read masses of documentation produced in foreign languages. It is also necessary to choose a system to which NEW languages can be rapidly added, i.e. a modular, interlingual one.

Much of the MT-related research performed in the US is being applied elsewhere. No nationwide project utilizing the best research talents in NLP has been attempted in the U.S. in over two decades. Today, Darpa is probably the only institution with the resources and scope to mount a large-scale MT effort successfully. Such an effort would harness and coordinate NLP work of various kinds and would create a setting in which new innovations could be used within this country first.

A second strategic reason pertains to interproject collaborations. Currently, there is relatively little collaboration and sharing of resources and expertise among NLP research groups in this country. A new national agenda with a set of clearly focused goals could serve as an integrating agent. The development of a standard interlingua representation, a set of standardized lexicons, one or more grammars, support tools and interfaces, and additional software, can shape much future NLP research in this country by enabling researchers to make use of existing work and tools with much less effort than is currently the case.

Technical Reasons for an MT Effort:

Steady developments in various aspects of NLP make available large portions of an MT system more or less off the shelf, which greatly facilitates the construction of new MT systems. These developments are the following:

1. Clearer understanding of semantics: Recent refinements of taxonomical ontologies of representation provide an interlingua-like basis for a new, more powerful, MT. Making maximal use of the high-level linguistic and semantic generalizations shared among languages, one can minimize language-to-language lexical or structural transfer rules and so increase the portability of the system across domains.
2. More complete grammars: Development of grammars is an ongoing process. There exist today grammars that cover English (and other languages such as German, Chinese, Japanese, and French) far more extensively than the most comprehensive grammars of 20 years ago did.
3. Better existing generation and parsing technology: Single-sentence parsing and generation has been studied to the point where a number of well-established paradigms and algorithms exist, each with known strengths and weaknesses, a situation which greatly facilitates the construction of a new MT system (in fact, in the last 5 years a number of general-purpose generators have been distributed: Penman, mumble, frege , etc.).
4. In addition, the number of existing MT systems and the amount of MT experience is also much larger than it was in the early days, especially in Europe and Japan.

An Interlingual Approach Versus Transfer Or Massive Statistics.

A fundamental technical notion in our proposal is interlinguality: it is one of the three basic structural methods for MT, contrasted with direct and transfer approaches. The direct method was used for early systems like SYSTRAN as well as large recent ones like SHALT from IBM Japan. If one is only every going to be interested in one language couple in one direction, as SYSTRAN originally was, there is no reason not to use it. We assume, however, that that is not our situation and multilinguality is essential. It should also be noted that some form of interlinguality is now becoming the standard position in AI-knowledge representation and our approach meshes best with that. The interlingua approach overcomes the problem of building thousands of transfer rules by using a central representation into which and from which all the languages are parsed and generated.

Of major concern is to design an interlingua which is both specific enough to allow simple and unambiguous processing and general enough to enable different approaches with different theoretical strengths to represent the information they can extract from the text. Fortunately, none of the parties involved have ever been committed to the highly formalized representation languages and systems which have been (and still are) popular in various areas of NLP, formalisms whose logical properties have been studied extensively but whose practical utility is low.

Consider now the following example:

"Mary was in a severe accident. She lost a foot."
vs. "Mary was buying cloth, but measured it incorrectly by accident. She lost a foot."

There is no statistical measure (e.g., no low-order n-grams) that will disambiguate reliably. Yet, if a sentence similar to the above concerned the Lybian Colonel or Abu Nidal it might be useful to have accurate intelligence. Language other than English have different ambiguities that must be resolved to translate to English or to fill a database for an analyst.

The interlingua approach is far better able to exploit domain knowledge in order to produce reliable translations than the other two approaches. The massive statistical approach is inimical to any infusion of domain knowledge or any comprehension of the language. Pure statistical translation had been rejected in the early years, but has been brought back to life in the recent IBM research effort. Experience has consistently shown that unaided statistical methods perform only at a low level which cannot be raised much, and only on a carefully selected materials (in the IBM project based on the copious high-quality parallel French-English Hansard texts from Canada -- data not found for other language pair. Even the 50 success claimed may depend crucially on order similarities between English and French. The paper claims that for 63 of tested sentences under 10 words, the most probable word order, based on trigram probabilities, was the correct one 80 of the time, which together produce the figure above.

The transfer approach is indeed capable of using domain knowledge, but the software engineering is much worse than an interlingual approach. If one must translate among N languages, there are N(N-1)/2 language pairs. A transfer approach would require on the order of N**2 transfer grammars (2 per language pair, one for each direction), if these must be augmented with domain semantics, a task that was gargantuan to start becomes totally intractable to hardiest of souls. In contrast, the interlingua approach requires 2N grammars (1 for analysis and 1 for generation for each language, into and out of the standardized common interlingual knowledge representation). Domain knowledge, though potentially complex, need be added only once per domain and retained in modular, reusable declarative data structures that serve as input to a unifying compiler. This compiler combines modular domain knowledge and grammar files dynamically to produce a run time translator among two languages for a given domain (or set of domains).

Statistics, although not the preferred translation paradigm, plays several important roles in MT, including: Once the meaning of a text is analyzed, selecting the most normative (frequent) rendition into words in each target language. Statistics can select collocations from large text corpora (such as the preferred use of ''pitch black'' rather than ''asphalt black''). Given a large potential lexicon, simple frequency analysis can direct the dictionary-building work towards the most frequent words first, so as to obtain maximal utility of a system during development phases. All evaluation metrics of fluency, accuracy and cost of translation are statistically based.

Machine translation systems must be concerned with the knowledge encoding, with modular software architectures, with good engineering, with scalable and evaluable systems development, much more so than with specific linguistic theories prevalent in modern transfer approaches. In practice, MT approaches motivated by theoretical-linguistic concerns, like EUROTRA, tend to be too driven by linguistic fashion (since their chief motivation is to be theoretically interesting rather than effective). This opinion is shared by the Japanese researchers. Thus, the 1989 JEIDA report concluded that linguistic theory had made no discernible contribution to the advance of MT. Key features of the cooperative approach we advocate are:

1. The use of an interlingua instead of transfer rules or statistical cooccurrences;
2. Modularity : both programs and data will be produced in a modular fashion allowing them to be assembled into a number of prototype MT systems;
3. Commitment to gradual increase in the levels of automation of the systems we create;
4. The central role of world knowledge in addition to knowledge about language;
5. The use of a representation based on commonsense semantics and pragmatics ;
6. Emphasis on the large scale of the systems under construction;
7. Ensuring portability across domains by building reusable tools and information repositories such as lexicons;
8. Developing a translator's workstation environment a) to facilitate the integration of the above modules and b) to support the creation of useful machine-aided translation systems at the earlier stages of the project, while the various automatic processing modules are being developed. Included here will be a separate, but compatible, lexicology workstation, to assist the incorporation of large-scale semantic, syntactic and collocational information from machine-readable dictionaries and text corpora.

The Modularity Assumption

Modularity is independent of interlinguality though opting for the latter requires the former. Strong modularity of language components would now be supported by most researchers and developers in MT, largely because it allows the addition of new languages with minimum dislocation. It is also essential if it is to be possible to treat different languages by different methods and to combine work at a range of sites. Agreeing on suitable interfaces is a practical not a theoretical matter, and the experience of EUROTRA has shown it is perfectly feasible (this is the main scientific contribution of EUROTRA).

In order to harness the NLP research potential in this country, a modular approach to the construction of prototype MT systsems is proposed. Under this approach, various sites will build various modules which can be assembled in various ways to construct various prototype systems.

Two advantages of the modular approach are: new languages and additional functionalities such as gisting can be added with minimal disruption to the existing system, and the system can support various theoretical approaches (which may be required by various languages, or which may be the best way to foster collaborations among groups with different research methodologies).

MT system modules are either theory-neutral or theory-based. Theory-neutral modules are typically receptacles of basic information, such as core lexicons, morphological information, models of the application domain, domain lexicons, etc. Theory-based modules are modules whose construction and performance depends on a particular theoretical approach (of Linguistics, semantics, etc.); typical instances are parsers, generators, and theory-based grammars.

In order to limit redundancy, this proposal calls for the straightforward incorporation of existing theory-neutral modules from any available source. Various lexicons and some theory-neutral grammars of several languages exist in the public domain. Only when such information is unavailable, or when the available information is not structured in a useful way, should a new module be constructed. In such cases, the proposal calls for the construction of a single module, to be shared by all participants in the MT program.

Some modules, however, must be structured to conform to the requirements of a particular theoretical approach. In order to allow various approaches to participate (and be tested) in the MT program, the proposal calls for the parallel construction of various theory-based modules that perform the same function. Different modules will be constructed at different sites, but the enforcement of an intermodule communication protocol will ensure that the modules are mutually replaceable.

To summarize the modularity issue, we propose to enforce standard interfaces, modular development and maintenance in the following dimensions:

- Modular knowledge bases
- Common "top" of ontology across all domains
  Combinable "subworld" ontologies for specific domains
  No language-specific info in knowledge bases, only domain info.
- Modular unification-grammar files
- High-level well-structured grammars for each language
  No domain-specific info in any language specification
- Unifying grammar compiler (or interpreter)
- Takes language, domain and dictionary to produce run-time working MT system No human needs to cope with the output object code of unifying compiler, in the same way that no one needs to look at output of ADA compiler once verified.

The advantages of this modular approach include the following:

1. Various projects and various theoretical approaches will be able to participate.
2. Projects need not have experience in all aspects of MT to participate.
3. Redundant development of modules will be eliminated.
4. Interproject collaboration will be stimulated throughout the U.S.
5. The common goal of translation will provide a more coherent focus for the various research endeavors and will facilitate the comparison of various approaches to tease out their strengths and weaknesses.
6. As new and promising projects are found, they can be included into the program.
7. The theory-neutral modules, all in a standard form, will be made available to the whole NLP community as a basic resource.
8. Large-scale lexicons, automatically constructed from text, can be used in parsing and generation and in interactive help.
9. Existing work on collocation, cooccurrence, and clustering of words and phrases can be put to use (for example, to guide lexicon construction).

The attached "mountain" figure shows a possible cooperation being established between the Center for Machine Translation (Carnegie Mellon University), the Computing Research Laboratory (New Mexico State University) and the Information Sciences Institute (University of Southern California), whose groups share both experience of MT and the above assumptions. Figure~ mountain shows the anticipated modules, starting from the bottom left-hand corner upward (parsing) going up, and then down again to the bottom right (generation). The approximate number of modules and the sites with expertise in them are indicated, with the sites responsible for the module in larger font. Boxes in the middle represent the tasks of knowledge acquisition and system integration, also annotated with responsible sites.

Gradual Improvement

It is important to note that not all modules will be required for the MAT system to run. A number of the more experimental aspects can be ''short-circuited'', resulting in a leaner representation of the input text (and weaker output, or correspondingly more work for the augmentor or post-editor). We plan to adopt the policy of first creating a machine-aided translation system and then gradually enhance the levels of automation in the subsequent versions by implementing new descriptive

In order to harness the NLP research potential in this country, a modular approach to the construction of prototype MT systems is proposed. Under this approach, various sites will build various modules which can be assembled in various ways to construct various prototype systems.

Two advantages of the modular approach are: new languages and additional functionalities such as gisting can be added with minimal disruption to the existing system, and the system can support various theoretical approaches (which may be required by various languages, or which may be the best way to foster collaborations among groups with different research methodologies).

MT system modules are either theory-neutral or theory-based. Theory-neutral modules are typically receptacles of basic information, such as core lexicons, morphological information, models of the application domain, domain lexicons, etc. Theory-based modules are modules whose construction and performance depends on a particular theoretical approach (of Linguistics, semantics, etc.); typical instances are parsers, generators, and theory-based grammars.

In order to limit redundancy, this proposal calls for the straightforward incorporation of existing theory-neutral modules from any available source. Various lexicons and some theory-neutral grammars of several languages exist in the public domain. Only when such information is unavailable, or when the available information is not structured in a useful way, should a new module be constructed. In such cases, the proposal calls for the construction of a single module, to be shared by all participants in the MT program.

Some modules, however, must be structured to conform to the requirements of a particular theoretical approach. In order to allow various approaches to participate (and be tested) in the MT program, the proposal calls for the parallel construction of various theory-based modules that perform the same function. Different modules will be constructed at different sites, but the enforcement of an intermodule communication protocol will ensure that the modules are mutually replaceable.

To summarize the modularity issue, we propose to enforce standard interfaces, modular development and maintenance in the following dimensions:

- Modular knowledge bases
- Common "top" of ontology across all domains
  Combinable "subworld" ontologies for specific domains
  No language-specific info in knowledge bases, only domain info.
- Modular unification-grammar files
- High-level well-structured grammars for each language
  No domain-specific info in any language specification
- Unifying grammar compiler (or interpreter)
- Takes language, domain and dictionary to produce run-time working MT system No human needs to cope with the output object code of unifying compiler, in the same way that no one needs to look at output of ADA compiler once verified.

The advantages of this modular approach include the following:

1. Various projects and various theoretical approaches will be able to participate.
2. Projects need not have experience in all aspects of MT to participate.
3. Redundant development of modules will be eliminated.
4. Interproject collaboration will be stimulated throughout the U.S.
5. The common goal of translation will provide a more coherent focus for the various research endeavors and will facilitate the comparison of various approaches to tease out their strengths and weaknesses.
6. As new and promising projects are found, they can be included into the program.
7. The theory-neutral modules, all in a standard form, will be made available to the whole NLP community as a basic resource.
8. Large-scale lexicons, automatically constructed from text, can be used in parsing and generation and in interactive help.
9. Existing work on collocation, cooccurrence, and clustering of words and phrases can be put to use (for example, to guide lexicon construction).

The attached "mountain" figure shows a possible cooperation being established between the Center for Machine Translation (Carnegie Mellon University), the Computing Research Laboratory (New Mexico State University) and the Information Sciences Institute (University of Southern California), whose groups share both experience of MT and the above assumptions. Figure~ mountain shows the anticipated modules, starting from the bottom left-hand corner upward (parsing) going up, and then down again to the bottom right (generation). The approximate number of modules and the sites with expertise in them are indicated, with the sites responsible for the module in larger font. Boxes in the middle represent the tasks of knowledge acquisition and system integration, also annotated with responsible sites.

Gradual Improvement

It is important to note that not all modules will be required for the MAT system to run. A number of the more experimental aspects can be "short-circuited", resulting in a leaner representation of the input text (and weaker output, or correspondingly more work for the augmentor or post-editor). We plan to adopt the policy of first creating a machine-aided translation system and then gradually enhance the levels of automation in the subsequent versions by implementing new descriptive

World Knowledge

Ours is an AI approach in that we shall, in processing expressions so as to select a particular interpretation, apply computationally expressed knowledge of the world, as well as our knowledge of language. We thus select the most sensible interpretation of ambiguous expressions, recovering the most sensible referents for pronouns and inferring information which is implicit. This knowledge of the world is general in the sense that we know a great deal about objects, actions, states, events and situations, such as the classes to which they belong and the attributes they possess. Through the application of such knowledge, we weed out incoherent interpretations as they develop and select the most appropriate interpretation from those that survive.

Commonsense Semantics and Pragmatics

A crucial component is a realistic pragmatics, bringing in the best of AI work on speech act, belief etc. phenomena. These are now tractable and usable notions in MT systems. We shall commit ourselves to commonsense semantic approaches rather than formal ones since these have not proved fruitful in MT in any language. This will also involve a commitment to algorithmic elements of AI-based semantics (such as Preference Semantics) that have already proved useful in message-understanding work, and have an intimate connection with understanding of ill-formed, metaphor-laden text that is the normal form of actual documents.

In order to build working, portable prototype systems, the most practical and useful notations must be used. As mentioned above, the selection of notations whose properties are desirable on formal grounds but whose practical utility is low will be avoided.

In order to produce MT of superior quality that existing systems, one of the most powerful key ideas is the use of discourse-related and pragmatic terms. Most MT systems operate on a sentence-by-sentence basis only; they take no account of the discourse structure. Given recent work on discourse structure at various centers in the U.S., structural information should be taken into account and can be used to improve the quality of the translation. Similarly, pragmatic information, such as Speech Acts, reference treatment, and perhaps even some stylistic notions (to the extent that notations have been developed to represent them) will be used to improve the quality of the translation.

Scale

We emphasize scale phenomena, both in the sense of bringing large-scale lexical material automatically via existing work on machine readable dictionaries, but also making use where possible of statistically-based work on corpora to guide lexical entry selection, corrigibility of sentences to particular syntax rules etc.

Portability

- Construct reusable tools, general information repositories (e.g., lexicons, grammars)
- Establish nationwide resources in standard form
- Ensure future reusability

- Construct reusable tools, general information repositories (e.g., lexicons, grammars)
- Establish nationwide resources in standard form
- Ensure future reusability

One of the well-known weaknesses of current MT systems is their limited applicability. In order to achieve an acceptable level of translation quality, the current brute-force approaches require large collections of translation rules which invariably contain increasingly domain-specific information. Porting these systems to a new domain becomes a major undertaking.

By using the newest NLP technology while focusing on the development and use of a number of very general information resources (such as a high-level concept ontology under which domain-specific ontologies are subordinated, and a general lexicon for closed-class words), this proposal is aimed at overcoming the problem of domain-dependence without compromising on translation quality.

A major factor supporting the domain-independence is the ability to acquire information --- conceptual, lexical, phrasal, translational --- interactively during the translation process. When the system encounters input it cannot handle, it queries the human assistant, who decides what type of information the input is and then inserts appropriate definitions into the system's information banks for future use, using the interfaces and acquisition tools provided. The proposed MT program devotes a large amount of effort on the development of interactive acquisition software and interfaces, via the notions of the Translator's and Lexicologist's workstations.

The strengths and weaknesses of this interlingual approach

The strengths of the interlingua approach have been briefly discussed above.

The central weakness is the necessity to build a knowledge base, and therefore the initial development cost, though it can be amortized over other languages, and many translations in the context of a well-engineered modular system.

We would like now to defend the interlingua approach against three most commonly held negative opinions.

OPINION 1: An interlingual approach forces unneeded processing

If a source language has, say, an expression which is three ways ambiguous and some target language has an expression which has precisely the same three-way ambiguity, unanalyzed why not simply carry the ambiguity from the source to the target and let the reader figure it out? Why disambiguate needlessly?

The response is, on the one hand, that a third language probably has different expression for each of the possible interpretations, so that if the same representational apparatus is to be applied to translations between the source language and a third language or from the target language and a third language, such processing is necessary in any case. On the other hand, a quick inspection of bilingual dictionaries shows that cases of complete correspondence of ambiguity across languages is extremely rare, even in closely related languages such as German and Dutch.

The issue of "since we sometimes can get away with less processing, why risk doing unnecessary work?" can be compared with intelligence-gathering work, where much of the effort is routine; information often confirms expectations; and therefore much of the work is "unnecessary." With such an attitude, all unexpected, important intelligence would often be ignored, much to the detriment of the analysts and policymakers. Ignoring meaning in translation because it need not always be interpreted, is an equally flawed philosophy. The times when deeper analysis is required can be absolutely crucial to produce meaningful, rather than misleading, translations.

OPINION 2: Interlingual approaches are heavily knowledge dependent and the task of working out appropriate representations is too demanding to be practical.

It has been our experience that some, even if incomplete, level of knowledge representation is crucial to machine translation. The need for such knowledge is especially obvious in the translation of technical text, where translations based on a general knowledge of the world are markedly inferior to translations based on a specific knowledge of the subject domain of the translation. Large-scale knowledge bases are being actively developed in the field, and domain models have come to be considered a standard component of many AI-related application systems. In our system, acquisition of world knowledge will be an ongoing task, and we fully intend to prove the feasibility of working with large knowledge bases in practical terms. To wit, EDR laboratories, and Fujitsu in Japan have come to the same conclusion and are actively building large knowledge-bases for interlingua-based machine translation, with initial success.

OPINION 3: interlingual approaches are based on a particular language, thus creating unnatural analyses for other languages.

This is the "cultural imperialism" argument. If, however, there exists such a thing as a universal descriptive linguistic framework, then there is no reason to assume that language imperialism must be a side-effect of the interlingual approach. Our experience in the development of an interlingual representation based on a cross-linguistic comparison of parallel texts has indicated, at least, that such a language independent framework is possible. But even if no such framework exists, then at worst, such language particular bias would simply be defect of the approach, rather than wholly invalidating it.

A standard example here would be the case of the verb "wear" in English and the problem of expressing the notion in Japanese or Chinese. It so happens that in Japanese the corresponding verb depends entirely on what is worn e.g. shoes (verb= hateiru), coat (verb= kiteiru), spectacles (verb= kaketeiru) and so on (and similarly for Chinese). It is thus reasonable to say that Japanese does not have a concept of "wear" in the way English does. However, that observation is no kind of argument at all against an interlingual approach, merely one for intelligent generation. In an interlingual environment there will be at least one interlingual node (which may or may not correspond to "wear") that links the relevant sense representations. The crucial point is that it would be the intelligent Japanese generator (since no problem arises in the Japanese to English direction) that makes the choice of output verb based simply on selection semantics (e.g. if the worn object is "koutoo" the verb is "kiteiru" and so on).

Conclusion

There are several aspects of the knowledge-based interlingua MT project that incur a measure of risk. Foremost among these is the distributed management risk among the three centers. Although it is clearly in the national interest to establish several sites developing MT technology in mutual cooperation, special effort must be made to address communication, establishment of standards, mutual responsibility relationships, fall-back positions and so on. We think this is an eminently manageable risk, but nonetheless a omnipresent one. We are fully cognizant of this risk, and are prepared to minimize it by establishing common procedures, open lines of communication, and accommodation where necessary.

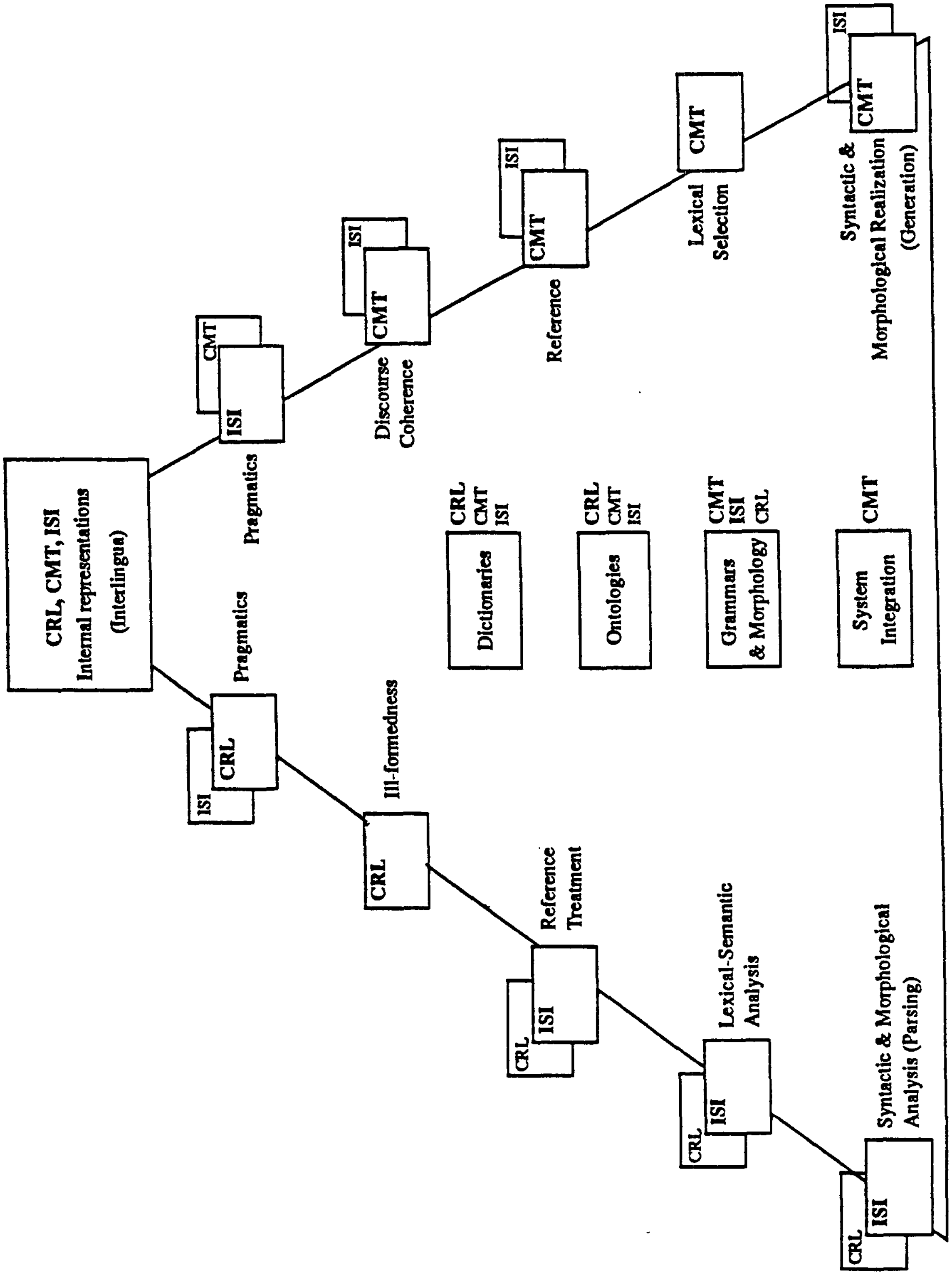

Figure 1: MAT System Modules.