# MACHINE TRANSLATION



Publication programme, 1993: Volume 8 (quarterly).
Institutional subscription prices, per volume: Dfl. 324,—/$ 202.50 including postage.
Individuals may subscribe at the reduced rate of Dfl. 162,—/$ 89.00 per volume. Members of the Association for Computational Linguistics: Dfl. 145,—/$ 74.00. They must declare that the subscription is for their own private use, it will not replace any institutional subscription, and it will not be put at the disposal of any library.
Subscriptions should be sent to **Kluwer Academic Publishers Group, P.O. Box 322, 3300 AH Dordrecht, The Netherlands,** or at **P.O. Box 358, Accord Station, Hingham, MA 02018-0358, U.S.A.,** or to any subscription agent. Private subscriptions should be sent direct to the publishers. Changes of mailing address should be notified together with our latest label.
For advertisement rates, prices of back volumes, and other information, apply to Kluwer Academic Publishers, P.O. Box 17, 3300 AA Dordrecht, The Netherlands.

Kluwer Academic Publishers incorporates the publishing programmes of D. Reidel, Martinus Nijhoff, Dr W. Junk and MTP Press.

MACHINE TRANSLATION

*Volume 7, No. 4 1992/1993*

FORUM ISSUE

*Current Research in Machine Translation*

Editor's Note 229

HAROLD L. SOMERS / Current Research in Machine Translation 231

DOUG ARNOLD / Reaction to Somers: Current Research in Machine Translation 247

LAURENT R. BOURBEAU / Current MT Research Orientation/Disorientation 253

JOHN HUTCHINS / And after the Second Generation... 259

PIERRE ISABELLE / Current Research in Machine Translation: A Reply to Somers 265

MARGARET KING / State of the Art and Perspectives 273

ERICH STEINER / Producers – Users – Customers: Towards a Differentiated Evaluation of Research in Machine Translation 281

KEH-YIH SU and JING-SHIN CHANG / Why MT Systems Are Still Not Widely Used 285

MURIEL VASCONCELLOS / What Do We Want from MT? 293

JOHN S. WHITE / Responses to Current Research in Machine Translation by Harold Somers 303

YORICK WILKS / Commentary on Somers' Article 309

HAROLD L. SOMERS / A Response to the Comments 315

The 'KLUWER' LaTeX Style File: Instructions for Authors 323

Keyword Index 331

Volume Contents 333

# What Do We Want from MT?

MURIEL VASCONCELLOS
*Machine Translation Consultant,*
*1739½ Corcoran Street, N.W.,*
*Washington D.C., 20009, U.S.A.*

## 1. The Goals of MT Research

As a long-time student of linguistics and a developer and user of machine translation, I feel moved to respond to Harold Somers' call for new architectures by asking the basic question: What is it that we want from MT? The definition of research priorities must necessarily flow from the purpose(s) sought. The needs will differ depending on what we hope to accomplish.

We must first decide whether or not it is a meaningful research goal to use MT merely as a test bed for the solution of challenging linguistic problems – in other words, to develop models in which the purpose is to come up with specific solutions rather than to build an integral translation system. Certainly most linguists would agree that linguistic problems can only be dealt with in the context of real language. Man-made scenarios with contrived input do not lead to solutions that stand up in a practical world. Systems must test pieces of real language, no matter how narrow the subset. Moreover, since their structural and lexical elements come from the larger language, the study of such pieces is worthwhile and valid only to the extent that the input is natural and unconstrained and has originally been generated for a clearly defined purpose. Personally, I also believe that the knowledge bases of such systems must incorporate necessary links to the rest of the larger language. In my view, research on selectively encoded toy systems that do not take these criteria into account runs two major risks: (1) it is not likely to stand up theoretically on larger corpora, and, therefore, (2) it is not likely to contribute to overall progress in machine translation.

If this position is accepted, and we agree that MT research should envision the development of working systems, then we must specify goals not for linguistic research as such but rather for integral *systems*. Several possible goals come to mind:

**PMT** To solve all the problems of translation, from analysis of the source text to generation of the target, on any subject, from and into any language in an efficient way, or what might be termed "perfect all-languages MT" (fully automatic high-quality MT with the added

*Machine Translation* **7:** 293–301, 1993.

capacity of multilingual reversibility), or $PMT_S$, in certain subsets of language combinations ("perfect subset-of-languages MT");

**GIMT** To generate reasonably correct translations in all subject areas from and into $n$ different languages based on an interlingua as pivot ("general interlingual MT"), or $GIMT_S$, even better translations using the same approach with certain subsets of language combinations ("general interlingual subset-of-languages MT");

**RIMT** To generate pretty good translations in restricted domains from and into $n$ different languages based on an interlingua as pivot ("restricted interlingual MT"), or $RIMT_S$, even better translations using the same approach with certain subsets of language combinations ("restricted interlingual subset-of-languages MT");

**GNIMT** To produce rough translations in any domain, from and into subsets of language combinations, such that they can be used as an information tool or as input for a postediting mode, with research focused on high-priority trouble spots that repeatedly impair the smooth functioning of the system in operational settings ("general-purpose non-interlingual MT" – I would dearly love to see an end to references to first, second, and third "generations"!), or

**RNIMT** To turn out much better translations using this approach in restricted domains ("restricted non-interlingual MT").

## 2. Bridging the Transitions

The majority of MT researchers would dismiss perfect MT (PMT and $PMT_S$) as unattainable, and Somers no doubt agrees. The question, then, is whether there is any point to holding this out as the ultimate goal. I happen to think that if it is out of the question to begin with, there is no reason to aspire to it. I do not believe it is productive to keep trying for the impossible in the hope of gradually pushing back the "frontier". I would go even further and say that to insist on perfect MT as a long-term goal is to waste time in a counterproductive distraction and to deprive research of the focus it urgently needs.

If this reasoning is followed and perfect MT is abandoned, then it also follows that there is no use striving for an architecture that bridges the gap from general interlingual MT to perfect MT (GIMT $\Rightarrow$ PMT), or for that matter from general-purpose non-interlingual MT to perfect MT (GNIMT $\Rightarrow$ PMT). To do so is to build a "bridge across forever".

This leaves general interlingual MT (GIMT or $GIMT_S$) as the only remote goal that we do not yet have any kind of a handle on. Restricted interlingual MT (at least $RIMT_S$) already exists, as do non-interlingual

general MT (GNIMT) and restricted MT (RNIMT), although in all cases there is considerable room for improvement and growth.

Here again, the transitions from one category to another are problematic. In the first case, it has been seriously questioned whether a domain-specific application (RIMT or RNIMT), regardless of the set of languages addressed, can be extended to general-purpose translation without a major breakdown in system performance (Shann 1987, among others). While it is true that new approaches – for example, cognitive modeling based on selected or extensive use of domain knowledge (artificial intelligence) and new interactive architectures – are already enabling systems to extend their coverage with less of a decline in performance, this is still a far cry from general translation, and there is serious doubt whether the summing-up of micro-worlds will eventually lead a real world in which an MT system can comfortably tackle texts from a broad range of sources.

Secondly, there is also a question whether system competence in a subset of languages ($RIMT_s$) can be effectively extended to all languages (RIMT). The jury is still out on whether it is economic to attempt to capture universals common to all languages, whether such universals are formulable, whether they even resemble what we think they are today, and indeed whether they exist at all.

Finally, as for going from non-interlingual MT to interlingual MT, it is generally held that an interlingual system has to be built up from scratch with its own architecture and that the approach cannot be grafted onto another existing system, especially a system of "second-generation design". To my knowledge, however, this contention has yet to be put to the test. It stands to reason that this would only be tried with a restricted system such as METEO. Certainly no one would think of taking a robust non-interlingual system such as LOGOS, METAL, or SYSTRAN and attempting to make the entire system interlingual.

Thus, the transitions from one step to the next are probably, for inherent reasons, not do-able, or only do-able with such great difficulty that they are not worth the effort.

## 3. The Problems of General-Purpose MT

Before focusing more specifically on Somers' proposals, I would like to go a little further into the problems that we face when we attempt to develop MT systems capable of dealing with anything that comes over the transom – certainly a legitimate goal that cannot be ruled out entirely if we bear in mind the needs of agencies that are likely to fund research in MT. The bottom line is that general-purpose MT faces many of the same problems that lead us to say that perfect MT is impossible.

To begin with, general text in large quantities is bound to contain variant spellings, misspellings, and new words that are not found in the

MT system's lexicon. For want of a nail, · · · the battle may well be lost. If a word is not found, it may be impossible to get to the underlying structure. When "gapping" routines are invoked, the resulting loss of information cannot fail to impair the quality of the translation, sometimes with fatal results.

Also, with general translation there is a fair proportion of input that is ill-formed, and the haste of word processing has added to this perennial problem. A real example of this problem:

(1) Doe-faces llamas members of the camel family and were domesticated in the Andean highlands of Peru some 4,000 to 5,000 years ago. [PAHO document[1]]

A human reader, and a human translator, can quickly get to the sense of this statement; the gift of human perception is such that we make adjustments and work around missing information. But a machine will be stymied.

Moreover, real text, even when it is not ill-formed, can be difficult for humans, let alone machines, to parse. Another real-life example:

(2) To find effective solutions, plans, programs, and services need to take into account the differences between the sexes. [PAHO document]

In addition, language itself is constantly changing: old words take on new meanings and functions (e.g. nouns become verbs, transitive verbs become intransitive, etc.) in response to the pressure of new concepts; nonce-formations are introduced for special effect (e.g. "-gate" from "Watergate" is often used by extension to refer to other scandalous situations); and new vocabulary is coined (e.g. "attrit" from "attrition", etc.). It may not be efficient to introduce all these innovations into a system's lexicon.[2]

Other aspects of language undergo evolution as well: syntactic structures change over time. An example is the increased acceptance of topicalization in English:

(3) "The women who've gotten these calls, every one says they feel they've been raped", Allen said. [*Washington Post*]
(4) In last night's fight the teenagers who jumped in all were black. [*Washington Post*]

While it is possible to teach a parser to deal with topicalization, the price may be the loss of useful logic relating to modification and conjunction, and hence valuable information about associated syntactic and semantic

[1] Document encountered in production translation at the Pan American Health Organization, an international agency based in Washington, D.C., where I directed MT development and use for 15 years.

[2] In my own experience, with MT, I have not found it worthwhile to capture nonce-formations or metaphoric meanings in the lexicon unless the author has used them repeatedly throughout a fairly long text.

features. If parsing constraints are relaxed to allow for all contingencies, parsers will no longer be sufficiently discriminating to pick up important context-sensitive cues and elicit the precise translations that have been programmed for a system's more usual fodder.

There may also be ambiguities in a text which are potentially resolvable with the use of cognitive models and knowledge bases but which require a level of detail, in both breadth and depth, that would be impractical to contemplate in a general system. For example:

(5) Traditionally, Potomac has presented a haven for the wealthy who live in custom-built homes nestled in the countryside behind white clapboard fences that are worth in some cases between $300,000 to $500,000. [*Washington Post*]

If we leave aside the fact that it is usually houses, rather than fences, that are made of clapboard (an example of semantically ill-formed input), disambiguation is theoretically quite easy: nowadays it's no problem to train a system to know that clapboard fences don't cost $300,000. The rub is that it's virtually impossible to anticipate an ambiguity of this sort.

Sometimes, moreover, real sentences can be structured in such a way that even the most sophisticated cognitive model would be hard put to cope with them:

(6) Although exact figures are not available, it is estimated that about one-third of all residential sales today involve a seller who pays less than the standard 6 percent commission, up from perhaps 10 percent or less five years ago. [*Washington Post*]

Yet another problem arises for general-purpose MT when the target language calls for explicit distinctions that do no exist in the source language and material must therefore be added in order to make the target translation acceptable. To cite a very simple example, reports on ovarian cancer in English do not specify that the subject is female, but this distinction would be required in a great many target languages, and the default male translation in Spanish, for example, becomes ludicrous. Such information can readily be built into a knowledge base, but many other distinctions and types of information demanded by a target language are more elusive, yet they may be essential to achieving coherence in the translation. Their solution may require knowledge unrelated to the subject at hand or else internal to the discourse context – as, for example, in Japanese, when information about the relationship between the writer and the addressee becomes essential in order to produce a faithful translation. Elsewhere I have claimed that information of this kind, when it is dependent on unrelated world knowledge or the discourse context of a particular text, may be inherently unformulable (Vasconcellos 1989).

Behind many of these problems lies the fact that language communicates on multiple channels at once (Jakobson, 1960, Halliday 1967-68, etc.), and some of these have to do with the purpose of the specific communication. No amount of connectionist theory or parallel processing will ever save the day when knowledge about very special worlds or about the current discourse context, unlike all other contexts, is needed in order to unlock the intended meaning and bring it to the fore.

Translators are accustomed to dealing with the demands imposed by shifting from the universe of one language to that of another, and often their seemingly arbitrary introduction of changes in structure are in fact motivated by such concerns. Somers' dismissive gibe that the translater sees "structure-preserving translation as a last resort" shows a failure to appreciate the real task of translation.

Indeed, the problems of general translation are so daunting that new MT architectures and approaches to basic research, if aimed at general interlingual MT (GIMT), may prove to be more frustrating than fruitful. They could turn out to be a disappointing waste of money. At the same time, it should not be forgotten that existing general-purpose commercial MT systems – most notably LOGOS, METAL, SYSTRAN, and TOVNA in the West plus a number of others in Japan, all of them descendants of the "second-generation design" which today incorporate sophisticated semantic rules – have already achieved a fair degree of functional success in that they are meeting their customers' needs, and in some cases pretty well. Their developers' main goals (in addition to developing new language pairs in order to attract more business) are to extend the lexicons and to overcome the pesky recurrent problems that keep their customers from being entirely satisfied – and that turn away potential new ones. Whether the solutions lie in a new paradigm or in a creative flexing of the old one is of less interest to these developers than whether or not they work. Thus, for the case of general-purpose systems, it may well be that Somers' 1988 prediction, which he now recants, was in fact correct and that the approach most likely to produce results is heavy investments in lexicon-building coupled with development of user-friendly environments and linguistically-based editing functions specifically designed for MT.[3]

## 4. The Place for New Approaches

Given the difficulty of transitioning from restricted to general-purpose MT, from a subset of languages to all languages, and from non-interlingual to interlingual, my comments on Somers' proposals will be focused on the

[3] This last approach has already proved to be highly successful in the case of EDITSYS, the semiautomated postediting software for SYSTRAN introduced at the U.S. Air Force some 15 years ago (Bostad 1987).

more narrow perspective of improving system performance within specific categories.

In other words, it seems reasonable to look at research designed to expand the base and improve the performance of restricted interlingual MT, but not with the goal of achieving general-purpose translation. The same is true of restricted non-interlingual MT if it was not a piece from a general system to begin with. At the same time, of course, it is interesting to broaden the subset of languages that a restricted interlingual system can handle – but not, I would hope, with the goal of handling *all* languages.

Restricted systems of all kinds might benefit from a connectionist approach and experiments in parallel processing. Indeed, it is in the area of restricted systems that we should heed Somers' call and look for "viable alternatives to the procedural algorithmic strictly-typed programming style", the "importation of AI techniques", the "better linguistic theories", the environments that allow for dialoguing, and the others that he proposes. These approaches can be counted on to produce systems that are interesting and effective in limited settings. Here I would agree that there is room for innovation and that Somers' proposals are on the mark. We must only ask whether, even in the best of all worlds, many of the strategies developed will eventually migrate to general-purpose MT.

Thus the question remains: What should we do to meet the challenge of improving the non-interlingual general-purpose systems? Create new ones? Certainly not without at the same time taking advantage of the immense lexicons and rule bases now available – the fruit of many years of work by large teams of linguists – in the systems that already exist. These systems must not be scrapped or disparaged, because there is nothing else to take their place. *They need to be improved within their current paradigm.* The competence of general-purpose systems needs to be expanded to cover more text-types, more domains, and more linguistic structures. Problems such as anaphora resolution, scoping and conjoining, ellipsis, anacolutha, fragments, and unexpressed inferences need to be addressed within the framework of a general system without reducing it to a restricted system, because they cannot be fully tested on small pieces of language. Thus, while it might be an interesting experiment to identify a small area within a general-purpose non-interlingual system and smarten it up with a domain-specific knowledge base, the result would become a restricted subsystem; it would no longer be a general system and the exercise would not bring us very much closer to the goal of building up general-purpose MT.

There is one area, however, in which both general-purpose and restricted systems could benefit from concentrated research. Far greater investments are needed in the study of discourse organization from the standpoint of cohesion (Halliday and Hasan 1967), theme and information structure (Halliday 1967-68, etc., Grimes 1975, Vasconcellos 1985, 1986a, 1986b, etc.), and coherence (Vasconcellos 1989), which give fabric to a text and

make it easier to grasp. The value of cohesion and coherence can be seen in a re-working of example 2 above. It will be recalled that the original version was difficult to follow:

(2) To find effective solutions, plans, programs, and services need to take into account the differences between the sexes.

With a few "fixes" to give it cohesion and coherence, it becomes understandable:

(2′) In order for solutions to be effective, the development of plans, programs and services needs to take into account the differences between the sexes.

I feel safe in saying that the success of future research in MT will hinge largely on increased understanding of how text works at the discourse level.

In the meantime, we have to accept that there is no big new paradigm sitting out there on silent haunches waiting to move in and turn MT around. Recently there has been a renewed interest in machine translation that no one would have predicted in 1988. Thanks to the growing importance of information in the Information Age, to the massive availability of scientific and technical data in electronic form, and to the pressure to track information from foreign sources, MT has become fashionable for the first time in the United States in more than 30 years and the U.S. Government has opened its pockets to support not only studies on MT (e.g. JTEC 1991) but basic research as well. If this good will is to continue, and if MT is to regain credibility, research must be focused on very specific targets that are attainable and that will yield clear results in the near term. It must shy away from any suggestion that the results will be directly extensible to general-purpose MT. This is crucially important for the future of the technology: MT has gone to the mat once, and it cannot afford to do so a second time.

## REFERENCES

Bostad, Dale A. (1987) "Machine Translation: The USAF Experience." *Proceedings of the 28th Annual Conference of the American Translators Association*, ed. Karl Kummer, Medford (NJ): Learned Information, 435–443.

Grimes, Joseph E. (1975) *The Thread of Discourse*. The Hague: Mouton.

Halliday, M.A.K. (1967-68) "Notes on Transitivity and Theme in English," *Journal of Linguistics* **3**: 37–81 and 199–244; 179–215.

Halliday, M.A.K., and Ruqaiya Hasan (1976) *Cohesion in English*. London: Longman.

Jakobson, Roman (1959) "Linguistics and Poetics," In: *Style in Language*, ed. Thomas A. Sebeok. Cambridge (MA): MIT Press; New York; Wiley, 350–377.

JTEC: Japanese Technology Information Center (1991) *JTEC Panel on Machine Translation*. Baltimore: Loyola College, JTEC.

Shann, Patrick (1987) "Machine Translation: A Problem of Linguistic Engineering or of Cognitive Modelling?" In: *Machine Translation Today*, ed. Margaret King. Edinburgh; Edinburgh University Press, 71–90.

Vasconcellos, Muriel (1985) "Theme and Focus: Cross-Language Comparison via Translations from Extended Discourse." Unpublished Ph.D dissertation, Georgetown University, Washington, DC.

Vasconcellos, Muriel (1986a) "Functional Considerations in the Postediting of Machine-translated Output: Dealing with V(S)O versus SVO." *Computers and Translation* **1**: 21–38.

Vasconcellos, Muriel (1986b) "Humor through the Listener's Voice: A Functional Model for the Capture of Humor in Translation." *Babel* **32**: 134–145.

Vasconcellos, Muriel (1989) "Cohesion and Coherence in the Presentation of Machine Translation Products." In: *Georgetown University Round Table on Languages and Linguistics 1989*. Washington, DC: Georgetown University Press, 89–105.