[*Language Industry Monitor* 11, September-October 1992]

Statistical Methods Gaining Ground

The increasing popularity of statistics-based methods in computational linguistics was much in evidence at the June TMI conference in Canada. As Tony Whitecomb reports, it is a significant trend which could pave the way for major breakthroughs in natural LANguage processing.

Montreal, Quebec – For the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation, held here June 25 to 27, conference organizers Pierre Isabelle and Elliot Macklovitch chose the theme "Empiricist versus Rationalist Methods in MT." It was a significant decision because it underlined the profound methodological shift which has gradually taken place within the MT research community over the last few years. While the "rationalists" have traditionally based MT on formal linguistics and a model-theoretic (or rule-based approach), the "empiricists" favor the use of large corpora of existing translations, using example-based, statistical, or connectionist methods.This year's TMI conference, previously a rationalist stronghold, was therefore a telling indication of just how far the empiricist newcomers have come. They dominated the conference in every respect: in the number of papers, in the number of proponents, and in the power and eloquence of their arguments.

Hail, empiricist conquerors

Surprising was the ease with which the rationalists seemed to accept defeat at the hands of the empiricists. While thinking in terms of "victory" is premature (even empiricists have not yet delivered the long awaited high-quality MT system), the bandwagon seemed to be rolling: many an MT rationalist, sensing the eventual defeat in his camp, openly declared himself to be – now, if not always having been – in fact an empiricist.

The striking difference in the atmosphere at TMI '92 can be best illustrated by comparing the presentations IBM's statistics-based MT research group gave at TMI with those it gave at the 1988 COLING conference (the big biannual gathering of computational linguists). In 1988, Della Pietra apologized at great length for not having used any linguistic methods at all. Now, his collaborator Robert Mercer was an invited speaker and the star of the conference. Instead of offering apologies, he boldly asserted that "rationalist methods in MT will be on the scrapheap five years from now," basing his argument on IBM's decades of research into speech recognition. As he explained, results in that area improved dramatically after the rationalist methods of the early 1970s (phonetic based forms and phonological rules) were replaced by probalistic methods (the use of so-called training corpora). Quoting Fred Jellinek, the leader of IBM's speech recognition research group, he said, "Every time I fire a linguist, my performance goes up." In contrast to the 1970s and early 1980s, when IBM researchers enjoyed unrivaled computer capacity, today everybody has plenty of computer power. As a prerequisite to performing processing-intensive corpora analysis, this, he maintained, will assure the eventual victory of empirical methods in MT.

Carnegie Mellon University's Jaime Carbonell responded by pointing out that speech recognition, as it stands today, could hardly be called a success. Mercer replied that "empiricist SR [speech recognition] is an incredible success in comparison to rationalist SR, and it will be the same for MT ." CMU's Sergei Nirenburg and Xerox PARC's Martin Kay ventured further critical remarks, the latter raising the issue of the extended separation of contextual information in written language ("long- distance effects"), a phenomenon which "apparently didn't matter in speech recognition," but

all of this did not amount to a defense of the rationalist MT position in face of Mercer's attack on it.

The rationalist empire did not strike back

Finally, those "conservatives" who might have hoped that the invited speaker for the rationalist camp, the British-American NLP guru Yorick Wilks, could turn the tide, must have come away disappointed. Acknowledging the wealth of effort and experience which has gone into Systran, he commented, "we in the conservative camp have to accept Systran as the champion" – and that is hardly inspirational if you have bet your intellectual stakes on Eurotra, Rosetta, or any of the other ambitious rationalist MT projects of the recent past. As Wilks pointed out, Systran is still better than IBM's statistics-based system. "And now that they have exhausted the potential of the statistics, the IBM guys are adding good old rationalist elements (syntax, vocabularies, semantics) to what they claim to be their magic stone soup." Wilks cautiously suggested that there might be a natural ceiling to the success of pure statistical methods and that the choice of the English-French pair might have made it easy ("I'd like to see what happens with English-German").

Wilks also pointed out that the Bar-Hillel knowledge problem of MT ("the box is in the pen") is still unsolved and that this remains an argument in favor of a rationalist approach. On this issue, Wilks was supported by David Farwell (likewise from New Mexico State University), who reminded us that a lot of information lies outside the NL text ("in the translator's head"). Wilks further characterized IBM's approach as "Systran without tears": "Systran is indeed the paradigm for the future. That means: a lot of drudgery for years to come." Commenting on the brawn-over-brains approach of IBM, Wilks added, "what they're doing at IBM is not MT: it's an MT factory."

Practically all active conference participants agreed that the most likely and promising approach to pursue in the future is a hybrid approach based on example-based, statistics-oriented, and corpora-supported work which is backed by generalized syntactic, lexical, or semantic knowledge. Various Japanese contributions as well as the most recent work by IBM's Mercer group (as reported at the conference by Peter Brown) validated this approach.

But where are the parallel corpora?...

In contrast to the overall appeal of the empiricist approach headed by IBM, there was a notable lack of progress in dealing with more diverse corpora, either bilingual or multilingual. Researchers from IBM (Mercer, Brown) and those from the Bell Labs (Church, Gale) are still basing their research on the Canadian Hansards (the Canadian Acts of Parliament), as IBM did five years ago, because the Hansards are one of the few currently available sources of parallel text today. National or international programs like the tei or other logistics efforts to obtain more parallel corpora in the near future were hardly mentioned at TMI '92. One researcher from MIT (Lynette Hirschman) remarked that the cost of obtaining and aligning parallel corpora was prohibiting the spread of the empiricist approach – a very honest admission from a researcher whom you might expect to justify the rationalist approach on the basis of principles. Mercer's response was that maybe we should try not to do everything at once but start with just one language pair, such as English-French. Nonetheless, a number of papers dealt with the question of how to align bilingual corpora and offered various methods.

And the translators?

At an MT conference that may prove to have been historic because it marked, after more than forty years of MT research, a transition from one fundamental paradigm to another, the participation of the professional translators – the ultimate empiricists – might have seemed desirable. Unfortunately, those that were invited or initially had registered had other priorities (Brian Harris and Claude Bédard), whereas those that were present (mostly students of translation) did not take an active part in the presentations or discussions. In their absence, eminence grise Martin Kay proved to be an excellent spokesman on their behalf. In his brief concluding comments, he alluded to typical examples of translation fragments that abound in live bitexts (i.e., Canadian Airlines' *In Flight magazine*) which defy attempts at codification

in bilingual dictionaries, thereby giving his tacit support to the empiricist camp.