[*Language Industry Monitor* 15, May-June 1993]

ARPA’s Deep Pockets

News of another round of us government funding for language processing is a timely occasion to take a closer look at ARPA.

The American National Science Foundation (NSF) and the US Defence Department’s Advanced Research Projects Agency (ARPA) recently announced a significant new joint initiative for funding research in language technology. Up to US$ 2 million per year for three years has been earmarked for co-funding innovative, multi-disciplinary research projects, which, in the wording of the proposal, “support the longterm goal of achieving effective, general, human-computer communication through the medium of human language… and accelerate progress in the development of the scientific and technical foundations of automatic human language processing by computer.” The NSF has hitherto primarily funded basic research in a wide range of scientific fields; this is its first incursion into a more application-oriented domain. ARPA, of course, has funded spoken language and NLPoriented research programs for many years. Just what is ARPA and why does it fund this kind of research?”

The origins of this enigmatic government agency lie in the dark days of the Vietnam War, when ARPA was established as an alternative (i.e., quick and dirty) procurement channel for military operations in South East Asia. ARPA later evolved into the research wing of the US Department of Defence and subsidized research in a wide range of disciplines, much of it without overt military application.

During the Reagan- Bush years, ARPA gained a ‘D’ (for Defense) to become DARPA, because it was felt that the program should have a clear military orientation. The Clinton-Gore administration has since reversed that decision.

Whatever direction the political winds happen to be blowing, ARPA has long had deep pockets for funding computer research. Established in the 1970s, the ARPAnet was the forbearer of to day’s Internet; it pioneered the packetswitching system as a way of ensuring uninterrupted communications in time of war. Later, in the 1980s, the massive Strategic Computing Plan (SCP) was to result in a new generation of autonomous and highly-automated land, sea, and air vehicles capable of complex, far-ranging reconnaissance and attack missions. To win support for this effort, ARPA pointed to important advances it was fostering in AI, such as expert systems with common sense and natural language understanding. However, the us congressional Office of Technology Assessment evaluated the SCP and wrote, “unlike the Manhattan Project or the Manned Moon Landing Mission, which were principally engineering projects, the success of the DARPA program requires basic scientific breakthroughs, neither the timing nor the nature of which can be predicted.” The scp nonetheless unleashed the great AI boom of the 1980s, with venture capitalists swooping down on this nascent technology with the hope of making a killing.ARPA spinoffs, such as Verity, developer of the information retrieval package Topic, Thinking Machines, the builder of massively parallel computers, APEX, Carnegie Group, Cognitive Systems, and went on to flourish.

Others fell by the wayside. An unfortunate side effect, however, was a kind of AI burnout, with dashed expectations and a backlash of anti-AI sentiments.

ARPA involvement in language processing dates back to 1971, when it launched a five year “speech understanding project.” The goals of this early program were well defined, but the program’s overall achievements were limited and without meaningful technical advancement. As William S. Meisel, an industry analyst points out, the project’s most successful system, developed by two Carnegie Mellon students, had virtually no linguistic capabilities. ARPA resumed funding with more realistic expectations in 1984, when it initiated the Spoken and Written Language Program. This wide-ranging program was gi ven the broad mandate to supply the us government with voice technology which could be used for interactive access to databases and 'for system control, and over the past decade it has been periodically modified and re-approved. Some two dozen companies, institutions, and organizations are currently participating in various ongoing projects, structured in the form of "cooperative competitions." These include the Message Understanding Conference (MUC) for evaluating NLP software, TIPSTER Text for information extracting and routing, and the Text Retrieval Conference (TREC), for information retrieval.

In 1988, ARPA launched a second program, Spoken Language Systems, a five-year program centered around a pseudo-application, called the Air Travel Information Service (ATIS). A basic set of data was defined for ATIS, namely the names of ten American cities (now fifty) and the major airlines, and the so-called "common task" has been to develop an interactive system for querying the ATIS database and essentially going through all the steps it would take to book a real flight. This calls for robust, speaker-independent speech recognition and a dialogue system to manage the interaction within this admittedly highly restricted domain.

A number of well-functioning prototypes of the ATIS application can now be found in the participating research labs.

With common tasks likeATIS, ARPA tries to break a kind of gridlock in research, where systems don't get better unless they are tested in applications but don't get used in applications because they are not good enough. The common tasks are an attempt to jump-start the process of developing applications by defining pseudo-applications which test systems' ability to deal with "real" data. In the domain of real applications, system design, data manipulation, and human interface issues, while perhaps not as theoretically challenging, are nonetheless vital practical considerations which substantially determine the overall usability of the underlying technology.

More recently, ARPA started an MT program, currently in the second of a three year traj ectory. Dragon Systems, IBM, and a joint team of three American university research groups are now developing prototype systems to be tested on texts from the financial domain. The innovative if not yet entirely proven statistics-based approach of the IBM team certainly satisfies ARPA's avowed goal of funding innovative science; it also reflects ARPA's engineering orientation.

ARPA is also funding two repositories for linguistic data. The Consortium for Lexical Research (cLR) received a threeyear grant in 1991 to "facilitate the traNSFer of lexical data and software." The CLR is situated and administered by the Computing Laboratory of New Mexico State University. More recently, the Linguistic Data Consortium (IDC) was established at the University of Pennsylvania. Among other things, the IDC collects all of the data gathered for the various ARPA projects. This includes the TIMIT speech corpus, ATIS, the Penn Treebank (an annotated corpus), MUC terrorist reports, the A TC speech corpus, and the TIPSTER/TREC corpus. Because this data is available to the greater research community, researchers outside of the ARP Afunded sphere can also test their systems on the same data, thereby having a standard reference point when discussing the performance of their systems. The IDC was given an initial two-year grant by ARPA with the expectation that it would thereafter support itself by means of membership and licensing fees. Within the past year, all of the various ARPA programs mentioned above have been combined under the umbrella of ARPA 's "Human Language Processing" strategy.

One of the most important features of the ARPA language programs has been the development of appropriate evaluation methodologies, the evolution of which can most clearly be traced in the Message Understanding Conferences. For MUC-1 in 1987, which was coordinated by the Naval Command,Control, and Ocean SurveillanceCenter'sRDT&E Division (NRaD), six groups tested a set of twelve messages (tactical naval operations reports on ship sightings and engagements), ten of which were distributed in advance. But there was no specific task (ie, a specific goal) and no evaluation procedure. MUC- 2, two years later, saw a larger message base being distributed together with a manual evaluation procedure. At the behest of the participants, a specific task was developed which consisted of filling in templates for the events mentioned in the messages. However, it appeared that a more rigorous scoring method was needed because of the variation in scores.

For Muc-3, held in 1991, a new domain was chosen — Latin American terrorist reports garnered from foreign news services — and an automated, interactive scoring system was introduced. In addition, the systems were required to discern between relevant and irrelevant messages. MUC- 5, which will be held in August 1993, will present more challenging test corpus, the same data as is used in TIPSTER/TREC programs, and have yet further refined evaluation criteria. Messages willbe in two languages, English and Japanese, and two domains, joint ventures and microelectronic chip fabrication.

NLP systems are notoriously difficult to evaluate, but the ARPA community has made great progress by starting with simple tasks in narrow domains. The ARPA evaluation methods have attracted considerable interest both inside and outside the ARPA community, with at least two organizations, IBM and AT&T Bell Labs, participating voluntarily in ARPA programs because higher management recognized the usefulness of the evaluations. One of ARPA's most lasting contributions to language processing may be its cultivation of methodologies for testing NLP systems.

While ARPA can point to the impressive performance of some of the ATIS prototypes and is obviously helping advance language processing technogies with its support, ARPA's strategy also as some shortcomings. One of the most frequently heard criticisms is that the yearly evaluation intervals are too short a period for real development. As one listener eloquently put it at ARPA's 1992 Speech and Natural Language Workshop, "the ing R D. frequent evaluations may be encouraging researchers to hug the shore rather than to try risky, long term research." In more blunt terms, researchers may be more inclined towards basic hacking to ensure that they do well in the evaluations than expend effort searching for elegant solutions. Another critisicm is that the evaluations seem to be driving and/or defining research, instead of the other way around. However, this is a typical chicken-and-egg question, and clearly ARPA needs some kind of solid ground on which to start. It could hardly just specify methods and research approaches and sit back and wait for applications to emerge.

On another level, ARPA seems a bit removed from the commercial action. George Doddington, ARPA's program manager for spoken language, perceives the common tasks as a way of prodding his colleagues into developing some commercially viable consumer applications. At the National Academy of Science Symposium on Speech, held in Irvine, California, in February, he teased his audience of highpowered speech researchers by "demonstrating" two commercially available toys, Speak 'n Spell and a "talking" doll, which are based on a speech synthesis chip he helped develop while at Texas Instruments. For Doddington, the toys represented a successful marriage of a given technology and an appropriate application. Doddington's own ideas aside, this begs a deeper question: What does a military agency know about the mass market? Is ARPA the right organization to be overseeing and evangelizing the commercialization of language technology? Its astronomically expensive tanks and jet planes are not remotely market-driven. If anything, they reflect a virtually permanent subsidy of the US military-industrial complex over the past thirty years. The CEC has its legions of high-priced marketing consultants. Does ARPA? Today, the commercial action in speech is in the call processing arena, where a number of suppliers are quietly introducing small vocabulary, speaker independent recognition modules, some for a variety of languages. It is also, of course, in the handheld market, where those pocket "translators" of largely Asian origin are being sold by the millions. Does ARPA know this?"

More fundamentally, there may also be the danger that the ARPA strategy is skewed in its balance between research and development, echoing the AI debacle of the 1980s. Doddington has been known to try to provoke discussion by saying that he believes that given a corpus of an infinite size it should be able to process any text using purely statistical methods. It is inherently an engineering approach, not surpnsing in view of Doddington's background in speech, and it exasperates the NLP people for whom raw number crunching is not enough. While researchers by definition will call for more research, the lack of much basic understand of human communication on level of pragmatics and discourse is beyond argument, and it is here where researchers developing prototypes keep banging their heads against the limits of what is feasible. As Europeans know, putting development funds where research is needed is a prescription for Eurotralike disappointment. The time- and resource-intensive work that still needs to be done in NLP does not fit neatly into little one year packets. However, ARPA's new alliance with the NSF could signify a new era in which support for basic NLP research is balanced with support for development efforts.

Naturally the final judgement of ARPA's efforts in the field of speech and natural language will await the moment when the technologies ARPA has cultivated have blossomed into commercial products. This has yet to happen. One of the biggest obstacles may be portability. For MUC-3, the participants required on average a full person-year to adapt their systems to the MUC-3 test suite. BB&N's Madeline Bates pointed out at the NAS Symposium that much of the substantial work required to obtain the kind of performance achieved by the ATIS systems is not portable to new applications; it represents painstaking hand tooling. As with all too many experimental NLP applications, these systems do not offer broad linguistic coverage — a severe limitation. In any case, commercial breakthroughs can be expected first in the speech arena, as speech recognition and speech synthesis are now fairly mature technologies. Long,time ARPA protege BB&N is said to be on the verge of introducing a commercial speech recognition system and is ironing out with ARPA some intellectual property issues.

In the long term, ARPA's efforts may also be felt indirectly. IBM, Apple, and Microsoft have all hired speech researchers from the Carnegie Mellon University, for years a recipient of substantial ARPA funding. As speech recognition products from these companies start appearing on the market, they will share a common ARPA pedigree.

In the perspective of Remko Scha, previously with Philips and BB&N, now a professor at the University of Amsterdam, " ARPA's great achievement has been to get people to focus on specific problems. It's challenged claims and gotten researchers to put their money where their mouth is. Maybe it is time, though, for ARPA to leave these groups alone for five years or so. Let some new ideas crop up. However, now I think the time is right for an ARPA-like program in Europe."