[*Language Industry Monitor* 16, July-August 1993]

Europe’s Linguistic Superhighway

Why Europe needs a massive dose of language technology.

Luxembourg – If all goes according to plan, next year will see the launch of a mammoth new initiative on the part of DG XIII/E of the Commission of the European Communities to develop and promote a “linguistic infrastructure” for Europe.Tentatively called Language and Technology 2000, the program would be a part of the next phase of Community, funding for technology and industry, the so-called Fourth Framework (1994-1998). Although the Commission and the CEC member states have funded various aspects of language processing for many years – and many of the arguments for justifying these efforts may sound familiar – the landscape has clearly changed since the heyday of Eurotra, with arguments now more sharply contoured and colored by an unbureaucratlc urgency.

To promote LT 2000, DG XIII/E (information industry and market, language processing) has hit the road. In. advance of this autumn’s budgetary deliberations by the Council of Ministers, Language Awareness Days are being held in all of the CEC member states. The road show kicked off in June of last year in Lisbon, Portugal, followed by Athens in September, London in January, Pont-à-Mousson (France) in February, Brussels and the Hague in May, and Bonn in June. The proceedings will be brought to a close by Language Awareness Days in Spain, Denmark, Italy, and Ireland over the next few months. The gatherings have been organized by DG XIII/E officials, together with non-profit organizations in the respective countries, and have provided a platform for the Commission to bring its message home to its constituency. During these Days, DG XIII/E has been supported by consultants and government officials, and the proceedings have been interspersed by local organizations presenting bona fide commercial applications of language technology to provide a taste of the state of the art.

At these meetings, the director of DG XIII/E, Frans de Bruïne, has been highlighting in broad strokes a strategy for implementing this program. De Bruïine wants to see the results of diverse CEC-funded programs, such as Eurotra, the Multilingual Action Plan (MLAP), Esprit, LRE, and Telematica, coalesce with fruits borne by multi-national Eureka programs and miscellaneous national programs and flower within the Fourth Framework into a panoply of applications supported by common linguistic tools and resources and bolstered by standards and quality metrics. For De Bruïne, a grand European linguistic technology infrastructure for all of Europe’s languages will provide a much, needed platform upon which industry, academia, and national governments can further build.

This attention to language processing is, of course, just one aspect of a broader thrust to get Europe’s high tech industries back on track. The laissez-faire attitude towards research and technical development in the 1970s and 80s has given way to a more active effort to define an industrial policy for Europe. An underlying sentiment is that Europe, due to a lack strategic policy-making, dropped the ball in two of the most important .technologies of the modern era, semiconductors and biotechnology. As a result, Europe has had to concede an insurmountable lead by the Americans and the Japanese in these fields. With Europe’s IT industry languishing, language processing is a singular opportunity for Europe to take the lead in a technology that is destined in the next century to be as ubiquitous as – indeed inseparable from – the computer itself. Technology policy is a fiercely-debated topic these days on both sides of the Atlantic. In Europe, a committee headed by Wisse Dekker, ex-Philips chief, recommended abandoning the Esprit, RACE (telecommunications), and JESSI (semiconductors) R& TD programs because they have not delivered the industrial goods. In America, Bill Clinton has been promising new government initiatives to get US high-tech back in high gear. This in turn has created the unusual sight of Europeans warning the Americans not to imitate their failed strategies.

But in this infornation society, information technology remains too vital to abandon to the mercy of market forces; it is, after all, an enabling technology, not an end unto itself. The CEC believes that European IT is saddled with structural disadvantages and it feels it must try to level the playing field. The Japanese can devote "colossal" amounts to R&TD because its industries are clustered around banks which participate directly in strategic developmemt decisions and shield enterprises against short-term ups and downs. These renowned "keiretsu" ensure a high level of cooperation and solidarity between Japanese firms and strengthen their competitive advantage. In the US, a bold entrepreneurial spirit still flourishes, driven by easy access to venture capital. Moreover, Europeans also suspect Americans of still waving the flag of national security as a pretext for maintaining technological supremacy, despite the Cold War myth having been punctured.

With Europe lacking Japanese-style Keiretsu or American-style venture capitalists, European IT industry does not have comparable access to the long-term and/or high-risk capital it needs to move forward. With the decline of large American research-oriented companies like IBM and DEC and Europe's big names – Bull, Olivetti, Siemens-Nixdorf, Philips – floundering as well, a single company – or country – may never again be in the position to fund independently the development of the major linguistic resources so sorely needed. The result is a stalemate, where, as the Danzin report, an external study, puts it, "industrialists and publishers are not keen to risk the investment of private capital. Users, for their part, are not generally aware of the advances from which they might benefit. The market is thus deficient on two counts: research is not inspiring industry to develop commercial products, and users are not creating enough demand for them." If the future of language technology in Europe requires (and deserves) government intervention, it can only be mobilized on a European level at this point, which means action on the part of the Commission. This calls for a well articulated policy to point the way forward and substantial funding to realize these goals.

As we all know, the Eurotra program is now perceived as having been too ambitious. The rosy prognostications of the 1970s and 1980s far exceeded what the technology could deliver, resulting in what the CEC acknowledges has been "disappointment and disenchantment" in language language technology. However, it is generally agreed that the Eurotra program, into which the CEC pumped an estimated ECU 40 million, made important progress in cultivating computational linguistics throughout Europe. Although the program may not have achieved its original goals, language industry proponents are clearly anxious for the knowledge and resources accrued in Eurotra and other programs not to be dissipated by subsequent neglect. This sentiment is particulary acute in Holland, which hosted three illustrious MT projects during the 1980s, Eurotra, DLT (BSO), and Rosetta (Philips), that brought forth a high concentration of language engineering talent, but the same can be said to varying degrees for other countries as well. European activities in the field of language processing of the past fifteen years have generated a momentum which must be maintained. At the moment, however, the credibility needed to obtain further funding is in short supply. As one CEC official explains, "the Commission is counting on the LRE projects to supply it with tangible successes, thereby negating to some degree the 'popular' perception that Eurotra was a failure and paving the way forward to the next phase."

Gone, then, are the days of huge monolithic projects involving more than a hundred researchers. Small and feasible is the current creed, but with a difference – much closer coordination of efforts, or "project clustering," as DG XIII/E official Jan Roukens calls it. De Bruïne believes it should be possible to find useful and effective applications of "imperfect" technology which can be used today. But he points out that current efforts to do just this are hampered by the lack of a common framework, too little coordination among the small teams involved, and too little direction. Other weaknesses he identifies are a US lead in the market and a field largely populated by small companies targeting only national markets. Europe has also been failing to meet the demand for manpower trained to exploit technological advances. Whereas Japan trains eighty thousand engineers per year, France and Germany supply only forty-one thousand, according to recent CEC estimates. The Danzin Report calls for more "linguistic engineers and technicians" to help bridge the Grand Canyon between theoretical research and applications.

An urgent economic incentive for bankrolling these efforts is the massive document-crunchingburden faced by information-richEurope. As Jan Roukens says, "documents rule the world." Frans de Bruïne points out that a modest increase in efficiency of just 1 to 1.5 % in the ECU650 billion that European government and industry spend annually on documentation production and management could represent at least ECU5 billion per year in increased producti vity. But a more pressing issue than simply efficiency – a double-edged sword in view of Europe's unemployrnent problems – is the retrieval and sharing of information and finding the necessary information among the growing paper mountains and swelling electronic archives. Roukens says jobs can be created by better matching supply and demand in the services sector. He believes that there is many billions of ECU's worth of translation work untapped due to faulty logistics. There is currently no sophisticated means of matching the right jobs with the right specialists.

Remember, too, that the CEC, with nine working languages, is itself one of the biggest paper-churners in Europe. More languages are on the way as well, with new members due to be coming on board this decade. 1995 or thereabouts will probably see EFTA countries Norway, Sweden, Finland, and Austria signing on, with Hungary, the Czech Republic, Slovakia, and Poland waiting in the wings. In the next century, the numbers could be further swelled by other ex-Eastern Block countries, the Balkan entities, and the young states of the CIS. As Joshua Fishman recently pointed out in *Language International*, the number of languages in the CEC could forseeably reach twenty-nine in the next century – a mind-boggling 812 language pairs. Translating and interpreting services already form the largest part of the CEC's administrative budget. That bill will only be getting bigger.

However, it is doubtful that DG XIII/E will ever be able to argue successfully for such an enormous effort on purely economic terms. Tangible returns on such investments are notoriously difficult to measure and in any case take years to accumulate. CEC officials acknowledge that industrial takeup of Esprit technology has not been encouraging, while sceptics complain the ECU 13 billion pumped into Airbus will never be "recovered." But in the field of language, the Commission cannot permit itself simply to sit back and let market forces call the shots. The linguistic infrastructure debate quickly spills over in the arena of social policy, where it becomes patently clear that if only on the principle of subsidiarity – thright job for the right level of government – the Commission must do something. One of the policy centerpieces of the current campaign is a report issues last year titled *Vers une Infrastructure Linguistique Européenne* (Towards a European Linguistic Infrastructure), now translated and widely circulated. An accessible and occasionally eloquent document, the Danzin Report argues in very clear terms that the issue is not whether the CEC should embark upon this massive undertaking – it has no choice but to – but how.

It is a matter of social policy that makes a "linguistic infrastructure" so important, argue its authors. Taking a historical perspective, they see emerging digital tech- nologies as having a far greater impact on our society than printing but warn that "only those languages which were printed became major vehicles of communication and thought." A central tenet of the argument for a linguistic infrastructure is that equality of linguistic access to information should be considered a basic social right for all European citizens; in other words, "government in your language."

The onward march to European union as defined in the Maastricht treaty therefore requires commensurate measures to ensure that the plurality of European culture is not trampled underfoot. This means careful preservation of national, ethnic, or regional identity – however you slice the pie – and that eventually boils down to language. As Edinburgh University's John Laver, one of the CEC's external consultants for LT 2000, puts it, "language is the most central articulation of a culture." The ten official languages of the CEC (the nine working languages plus Irish Gaelic) are just part of the equation. There is also another ten or so indigenous languages in Europe, such as Welch, Basque, Catalan, Gallacian, and Frisian, which also deserve equal technological support for social reasons. Finally, there are also the languages spoken by Europe's immigrant population, notably Arabic and Turkish. Discriminate against a language – technologically speaking – and you discriminate against the people who speak it. Europeans need look no further than the ongoing horrors of the Balkan tragedy for a grim reminder of the dangers of ethnic isolationism and divisiveness.

If the development of a linguistic infrastructure for Europe was resigned to the whims of market forces, a dangerous imbalance would most likely arise between languages, with the scales tipped heavily in the favor of English. As the Danzin team argues, "the market is not a satisfactory regulator of language matters." Roukens has graphed by application language the software programs listed in the CECfunded Language Engineering Directory compiled by INK Luxembourg. Not surprisingly, English far outdistances the rest with 586 programs. French numbers 404 and German weighs in at 328. Danish and Greek trail with respectively 90 and 56. The Community, warns the Danzin Report, cannot allow a two-tier language situation to evolve. The report suggests there is an economic threshold at around forty to fifty million speakers; below that, the financial returns may not be high enough to justify the investment in purely commercial terms. If this is true and they miss the boat technologically, will Dutch, Danish, and Greek, for example, devolve into quaint patois?

Luxembourg's *glossocrats* therefore feel the heat of the Maastricht treaty on their backsides; they also espy stirrings in the language processing market. As John Laver points out, Frost and Sullivan's 1990 market research projections show several language technologies, most notably those in the speech processing arena, having reached or soon reaching critical mass in various European countries – US$50 million in annual sales per country. Laver's conclusion is an urgent one: "The debate about how to devise administrative mechanisms is subsidiary to the shape of a comprehensive policy on language and technology and should not be allowed to dominate the central question. If this opportunity is missed, it will not recur for most of the decade, by which time many of the questions that should have been addressed by conscious and careful policy-making will have been settled by default – probably with the help of language technology marketed by Europe's competitors."

While the precise level of funding for LT 2000 within the Fourth Framework has yet to be decided, De Bruïne is confident that it will be both "substantial and sufficient." The Danzin report recommends the staggering amount of 850 million ECUs for a four-year period. That works out to ECU 20 million per language per year, or, calculated differently, a hundred man-years per member state. That is a whole lot more than was pumped into Eurotra (an estimated ECU 40 million) and dwarfs the modest LRE outlay (ECU 22.5 for 1991-1994). Danzin is probably too hopeful; actual levels are likely to be in the order of two to three hundred million ECUs. For comparison sake, Japanese outlays for language processing research are on the order of 30 to 40 million ECUs annually, most notably for the ATR project. The massive nine-year Japanese Electronic Dictionary Research had been funded to the tune of ECU 90 million. But while the Japanese focus primarily on Japanese and English, Europe is faced with nine-plus languages. To place such deliberations in a wider context, there are now more Europeans working behind wordprocessors than toiling in the fields, yet European agriculture subsidies – an outlay with no returns – are in order of tens of billions of ECUs.

Of the ECU 13 billion proposed for research in the Fourth Framework, the next phase of CEC funding, 3.9 billion has been earmarked for information and communication technology. Will Europe's Council of Ministers be willing to devote a healthy chunk of this towards the development of linguistic infrastructure? DG XIII/E officials have compelling arguments, but will they get the funding they need to realize these grand aims? The answer is anyone's guess. To take a pessimistic point of view, DG XIII/E has a long, hard sell ahead of it and precious little previous success to show. The first round of LRE projects, though promising, is just getting off to a start and will not show tangible results for several years. Danzin authors tacitly acknowledge this problem, stating that it is imperative that any such outlay show prompt results. “There is no argument which would justify a large financial investment simply on the promise of a result if it only bore brilliant fruit some fifteen to twenty years later,” they write. This means, for example, improving Systran (widely used at the Commission) and getting industry involved right from the start.

In the past, CEC policy was to leave near-market research and development to industry, to ensure that their offerings remained “competitive,” and to fund only resolutely “pre-competitive” research. But as ex-Economist writer John Browning points out, “the problem with keeping research a safe distance from the marketplace is that it remains a safe distance from the market.” Is the Commission abandoning its pre-competitive stance? “Not exactly,” says Roukens. “Rather, we’re moving the focus of R&D much closer to the end-user. Computers have evolved so quickly in terms of the complexity of their actions and interactions, particularly with regard to the human-computer interface. The more complex the system, the more paramount – and difficult – it becomes to design effective human interfaces.” The long reiterative testing and evaluation process of building systems for working environments can also be considered pre-competitive. Such efforts may be closer to the market, yet they remain in a sense pre-competitive insofar as every company is forced to perform them. “When ideas have been implemented as working prototypes and potential users see something useful and are satisfied, that’s where we draw the line,” explains Roukens. A salient example of the new way of thinking is the Esprit Translator’s Workbench project.

Whatever the final form and size of LT 2000 within the Fourth Framework, evaluation of emerging speech and NLP technologies – a complex challenge – deserves to play a vital role. While DG XIII/E documents routinely allude to evaluation (usually in the same breath as standards), evaluation has yet to be given the central position it warrants. Suppliers – whether researchers or developers – will never be able sell their systems to government and industry without some objective measurement of their systems’ performance and capabilities, while government and industry will not be able to justify the large investment of language processing systems within their organizations without some reliable reference points. Whether DG XIII/E adopts the ARPA strategy of cooperative competitions or adopts another methodology, evaluation issues will surely be a decisive factor in the success of the program as a whole.

The Commission will also have to exert itself in evangelizing and promoting short, and long-term goals, giving the infonnation age workers of Europe some idea of what is in the pipeline. Current public perceptions of NLP are largely colored by egregious examples of machine translation or vague reports of automatic telephone interpreters – poorly exploited technology and dubious science fiction, in other words. For better or worse, AI has entered the public consciousness. NLP has not. Part of the problem is the still vast chasm between the research world and the real world, partly due to structural reasons. Researchers are judged by their scientific publications and forced to defend their tiny bits of turf with ever more obscure argots. The results are too often presented in an indigestible fashion, only comprehensible to a tiny, high-priesthood, and are almost never interpreted for wider consumption. Because researchers – rightfully or not – are not expected to think about such things as price tags, copyrights, and file formats, much research seems correspondingly irrelevant to the business world. Commercial operators offering today’s language technology frequently show a startling disdain for current research efforts. 1989 saw DG XIII/E launch a brief but imaginative publicity campaign within the framework of the Language Industries Survey. Is it time for a sustained effort of more of the same?