TOWARDS A EUROPEAN LANGUAGE INFRASTRUCTURE

Report

by A. DANZIN

and

the Strategic Planning Study Group

for

the Commission of the European Communities

(DG XIII)

31 March 1992

S T R U C T U R E O F T H E R E P O R T

C O N T E N T S

SUMMARY

For many years now the Commission has been carrying out work on the languages used in the Community. In September 1991 it commissioned a Study Group of outside experts to prepare a review of the current position regarding the automatic handling of mother tongues and to suggest a policy for the future. The following report, "Towards a European language infrastructure", is the fruit of its deliberations.

The report reaches conclusions of major significance.

These conclusions are prompted by the realization that there are three forces currently at work to modify the use of natural languages and their interactions:

- The impact of the transition from a locally-focused industrial age to an age of communication, knowledge and intelligence which spans the world. This impact brings profound changes in the handling of information flows and stocks which have become huge and unstable.

- The impact of the new concepts and products spawned by technological advance. Millions of new words have come into being in the last 50 years and new ones are being coined all the time. Requirements, on a scale hitherto unrealized, are arising for unambiguous language which will allow the complexities of science and technology, the key to economic success, to be encompassed without error.

- The impact of the new information technologies, specifically computer technology, on the use of mother tongues as a result of the transition from the "paper-based culture" to the "computer age". This impact is far greater, more sudden, more revolutionary in effect than any other previous development, notably printing. It can only be compared to the progression in human evolution from speech to the written word.

For the peoples of Europe these three forces represent a major new factor in the environment of their mother tongues. The response may be positive in the form of a technologically-driven adaptation of a language, the scale of which should not be underestimated. Or it may be negative in that the language is reluctant to adapt. In the latter case that language will cease to be a vehicle for scientific and technical thought and thus for economic and cultural thought. That is the view of the experts.

Europe post-Maastricht must take steps to resolve this problem if it is to preserve its diversity and be successful in economic and social terms. The objective is not simply to overcome the obstacles to communication between citizens of a single union, but to give everyone the opportunity of having a mother tongue which is equipped to meet the challenges of the 21st century.

One solution would be to let things take their own course. Looking at the consequences of the forces now at work on the language market, the Study Group concluded that there was no spontaneous balance. The normal outcome would be that a very small number of languages, perhaps just one language, would profit whilst the others would decline. If we want to avoid serious inequalities in the performance of the mother tongues of the Twelve and the social tensions that would entail, we need to have a consciously designed policy centred on a marriage of languages and computer science. The instrument of that policy is "language engineering". Its objective is the promotion of what we conventionally term the "language industries".

Once that conclusion is accepted, we are on familiar ground as regards the channels through which the Community can act. Those involved at all levels of innovation must be encouraged and persuaded to cooperate. The Community must aim to do more than just support research and development; it must also coordinate national measures, which will do more than anything else to ensure the survival of mother tongues, the training of specialists and users, the collection of basic linguistic resources, the launching of pilot projects and aid to pioneering entrepreneurs.

Special measures must also help those languages which are less widely used and thus offer a restricted market for applications.

The Commission must limit its role to that of initiator, coordinator and regulator (standards, rules on intellectual property, etc.). It must be aware of its rightful role as main developer and delegate the role of subcontractor, as far as possible in line with the principle of subsidiarity. This will culminate in a revision of structures to ensure that the measures taken are continuous over the period of time envisaged, and the report makes a number of suggestions to that end.

The scale of the work to be done must not be underestimated. If the critical threshold is not attained, the objectives will not be either. Though not a "material" venture, the promotion of language engineering will require sizeable resources to the tune of thousands of man-years for the Europe of the Twelve. The Study Group has estimated the level of funding needed for a policy to equip the Community with a satisfactory infrastructure. For the next five years it proposes a programme to a total value of 850 million ECU to be funded by the Commission, independently of specialized technology initiatives pursued under other programmes such as IMPACT and ESPRIT and over and above national measures.

In conclusion, the authors of the report point to the importance of their findings. Europe's future will depend on her ability to play a meaningful part in the "non-material" age. Mother tongues suitably modernized by computer methods will be the basic instruments of that new age. When submitting this report to the Commission the Study Group Chairman A. Danzin stated his conviction that the linguistic challenge had to be seen in economic and social terms as a phenomenon every bit as important as the emergence in the sixties of microelectronics and the computer software industries.

INTRODUCTION

Natural language and the information age

Western Europe, together with the USA and Japan, has entered upon the age of information. More than two thirds of the labour force, or some 100 million people in the European Community, work using information symbols (letters, figures, graphs) which relate to objects or services with which they have no physical contact. Staff using word processing systems in government departments, industry and commerce today outnumber the entire workforce employed in agriculture. That shows how far information-based operations have become central to the efficiency of systems for the creation and distribution of wealth.

If one can speak of an economic war, that is to say of competition to export unemployment from the strongest to the weakest, those strengths and weaknesses will be determined by the quality and quantity of transfers of information and knowledge.

Given this situation, the Europe of the single market needs to function as a high-performance information system, homogeneous in every respect. Despite differences in size between enterprises, the remoteness of peripheral regions and the peculiarities of natural language there must be equality of opportunity in the acquisition and exchange of knowledge. The language barrier is one of the prime obstacles to this homogeneity.

Following Maastricht, the Community will be working towards a Union. That goal can only be attained if it has long-term popular support. No region, no Member State, must be prompted to repudiate Europe and become inward-looking because of an inadequate information capacity. And the temptations of nationalism are directly linked to linguistic identity, which draws frontiers in the mind of the people even more firmly than political frontiers. We are seeing that at the moment in Eastern Europe and the Balkans. We must, therefore, protect the linguistic diversity which no one wishes to lose, but at the same time ensure that that diversity does not present an obstacle to the attainment of Union.

Japan and the United States face no obstacle of this kind. The language problem is central to the building of Europe, to its economic effectiveness and political stability.

Another and very different factor also focuses our attention on the language phenomenon. This is the impact of the new technologies on natural language. Likewise the new forms in which information has to be captured, processed, transmitted, stored and presented in order to be useful to science and technology and to benefit industrial competitiveness and administrative efficiency, particularly that of government departments.

Thus, at the very time when the language problem is assuming critical importance politically, the pressure of technology is transforming the way in which natural language is evolving. This review by the Study Group demonstrates that this convergence of technological change with the need for linguistic change offers great opportunities and calls for a suitable strategy.

Technology and the evolution of natural language

Everything indicates that Man was made to think and communicate.

Consequently the pattern of biological evolution whereby Man acquired the power of articulate speech must be seen as a phenomenon every bit as important as the development of the hand in direct relation to the growing sophistication of the brain. Of the inheritance passed down by generations of *homo sapiens*, the acquisition of language and the further refinement and precision of the spoken and written word are some of the most precious instruments of civilization. Words and the linking of them by syntax enable us to communicate and memorize. They are also the indispensable tools of thought. They enable us to speculate about the past, present and future and thus prodigiously increase the depth of our field of temporal perception. By virtue of that fact they become the instruments of power.

We do not know how and by what stages our ancestors moved on from mimicry and cries to a language rich in every manner of expressing emotions and desires, after giving names to things and people.

But we do know that present-day languages have a complex history, that they spring from a number of sources and have, like all areas of evolution, been subject to perpetual change and selective influences. *Like everything in the universe, languages are born, they live and die as companion pieces to the civilizations of which they are a part.*

Technology is not something separate from linguistic development. One specific technique, writing, was invented at a given moment in the history of spoken language, just a few thousand years ago. The only languages which survived into the future were those written down in stone or on shells. A new technical era began with the invention of the pen and parchment, or papyrus. This was a new threshold, which some languages of the time crossed more successfully than others, because cursive script, being joined up, is faster and does not have the same form as pictograms or symbols imprinted by stamping. These technical challenges played their part in the linguistic battle for predominance. They generally took place at the same time as other economic, political or religious events, and this explains why certain languages were dominant for a period or some languages disappeared.

A major development was clearly the invention of printing in China in about the tenth century A.D., which really came into its own four centuries later with Gutenberg. Without printing, such phenomena as the Age of Enlightenment, the French Revolution or the dawn of the industrial era would be unthinkable. Not all languages got into print. Those which did not benefit from the new technology then declined to dialect status. Only those languages which were printed became major vehicles of communication and thought. But printing has not caused any fundamental changes in those languages which benefitted from it.

Because of printing, people today who were born before 1950 are part of the paper-based culture, unlike their children, born after 1970, who are part of the computer age where things are rapidly changing.

The growth of computer technology

Our era is in fact experiencing a new stage in these successive transformations of written and spoken language and one which is far more important than Gutenberg because language is not only being enriched but actively changed and adapted to a variety of requirements by the automatic processing of sounds and symbols made possible by electronics and computer science. Proof of this is provided by a few examples of conversion or analysis work done by machines.

Optical scanning enables characters and words to be recognized after printing or typing. Machines will soon be able to recognize handwriting too. Intelligible sounds can already be generated by automated scanning. Machines will read and talk, thus converting the written word into speech. Progress is being made the other way round too. Henceforth a computer can have ears and recognize a limited series of simple isolated words.

But progress is so fast that it is possible, by the beginning of the next century, to conceive of systems which can understand speech and dictating machines which we shall speak into and which will then return a written text. Enormous efforts are needed in phonetics in conjunction with linguistics and appropriate software, but there is no reason why the ambitions should not be realized and we may see huge advances in document handling, in administrative work, the control of robots and household appliances, the operation of transport systems and vehicles, devices to correct certain sensory or motor handicaps, safety mechanisms, etc.

Word processors are now commonplace. They have transformed secretarial work but are still in their infancy. Some are already able to correct spelling mistakes caused by semantic or syntax error.

Others operate only with a closely defined vocabulary and controlled syntax. These produce texts which are free of the ambiguities of everyday language and are used in simplified machine translation systems.

We know that computers can scan text, give rapid word lists and frequency counts, identify specific syntactic forms, search for keywords and so provide a searcher almost instantaneously with the bibliographical references he needs from a huge mass of documents. By analysing the content of documents, machines can also help with filing and communication of the information retrieved to predesignated individuals on-line. Libraries will no longer be repositories of books to be browsed through but immense intelligent memories which can be remotely accessed.

Machine translation of written texts is still in its infancy but it is already possible, with certain language pairs, to scan an article in a specialized journal quickly to see if it warrants more careful translation. There are "pocket translators" to help businessmen understand the gist of their business correspondence and these are sometimes adequate for their purpose. Assembly and maintenance instructions for items manufactured by multinationals are increasingly published in several languages by computer systems which can check text as it is read in to ensure that it can be readily recognized by the translation software. And it is not impossible, by the next century, to imagine telephone links between callers who do not know each other's languages but are assisted by an electronic operator providing simultaneous translation. The big Japanese project on this has already demonstrated a simple conversation between a Japanese and an English speaker. The automatic operator will have to be complex, rapid and extremely well versed in the peculiar features of each langauge. It is likely, therefore, that only a limited number of language pairs will be able to benefit.

The aforegoing examples are not an exhaustive list of computer applications which can help our senses and brain in improving the use of our mother tongues and conversions between them. But they are enough to demonstrate the considerable impact of the new technologies on language tools. *That impact is far greater and far more sudden, and its repercussions far wider than the impact made by printing. We must recognize the fact and realize that this phenomenon will have some of the most important consequences of our time.*

The transformation of language requirements

The richness of mother tongues, with all their nuances, ambiguities and their ability to stimulate the imagination and emotions, is as essential to human social life as it ever was. Poetry, affection, love, are all communicated by the play of words and the way they are put together. Diplomatic negotiations or trade talks make use of the flexibility of language. There could be no excuse for subjecting the accumulated culture and wit of our mother tongues to mutilation by machines.

But side by side with this literary language we see a radically different area emerging and developing. There are powerful pressures, economic in origin, which favour the use of a reduced language from which all ambiguities have been removed. Whatever the area of technology in question, the information held in databases and banks and in large optoelectronic memories used by the various professions has to be stored in accordance with strictly defined rules. It must be proof gainst any errors of interpretation. It uses simplified syntax and vocabulary.

The same need to remove all doubts and keep to vocabularies and forms involving some degree of machine handling is reflected in the preparation of manuals of instruction for the use, installation and maintenance of appliances manufactured by industry for use in the professions or in the home.

Supply of an Airbus, a weapons system or space launcher system, a nuclear power station, telecommunications network or "turn-key" industrial plant must necessarily be accompanied by documentation which in paper terms may weigh anything from one to several tonnes. Not a single word of this huge body of information must lend itself to misinterpretation. In scientific and technical research, results must be communicable in such a way that users can consult them remotely in a rapid and reliable manner, and this requires special forms of preparation and editing. This trend is gaining ground in areas such as law and medicine as these become more and more complex.

It is in these areas of information essential to economic activity that electronic language handling systems have a part to play. There is no need to look unnecessarily for new applications for computer hardware and software; what we have to do is find an answer to the vital requirements of reliability in the complex world of science and technology and their industrial and commercial applications.

These evolutionary pressures on what we might term the "language of technology" are exacerbated by the globalization of the economy, the wider movement of goods, services, capital and people over vast geographical distances. A response is also needed to the neologisms which proliferate as a result of the need to find names for concepts, elementary or complex products and uses which did not previously exist. This expansion in nomenclatures is enormous. Over the last forty years, several million specific terms relating to new products or activities have appeared. The seat of this terminological explosion lies in a small number of countries and thus a small number of languages. The introduction of these new words into other languages has to be properly overseen.

The immensity of the task and the need for reliability means that computers have to be used. The same goes for the hypertexts and hypermedia created in response to the needs of business, industry and education.

Emergence and development of the language industries

In the evolution of language, then, a new pressure is necessarily being brought to bear: the need for new instruments for monolingual and multilingual processing. This pressure has led to the emergence and development of a new area of business and industry which is coming to be known as the "language industries" based on "language engineering".

These terms are surprising in their novelty. Hitherto, anything to do with natural language was considered as readily available and free, like the air we breathe or the water we drink (or used to). We have to accept that that age is past, and the fact will raise serious questions in connection with intellectual and industrial property. In future, specialist companies will offer "language tools" in the same way as computer companies currently market software. This development and sale of language products is a hallmark of the "non-material age" on which we have embarked without realizing it. The current phase is exemplified by a wide scatter of resources, poor matching of supply and demand and a relatively low turnover on the part of the specialist companies concerned. But the modest and diffident manner in which the language industries are entering the market must not lead us to underestimate their importance. It creates a new difficulty: the real winners are not the product developers but the users, to whom the continuing predominance of their mother tongue is a major advantage. This shift in the focus of benefits away from the product developer and towards the user offers no incentive to pioneering entrepreneurs. Barring exceptions, product developers will have to wait a long time before they see any return on a pioneering venture.

The scope of the language industries

The language industries should be taken hereafter as meaning all technical and economic activities concerned with the provision of these new language tools: basic linguistic resources, means of conversion between spoken and written language, semantic and syntactic processing, conversions from one language to another, text and speech analysis and all other technological procedures applied to the use of natural language. More generally, everything which modifies written and spoken language via the use of new technologies of any kind. This new language engineering is essentially geared to its beneficiaries who are persons engaged in business and industry and government and private administration.

This definition excludes matters concerned with information processing, transmission and storage using what is commonly referred to as the "information and communication technologies", or ICTs. Also excluded are computer programming languages as such.

There are, however, common areas which form interfaces between the language industries proper, ICT applications notably in the rich seams of knowledge represented by the databases and banks, and what is generally termed "communications", including the mass media. It goes without saying that the policies proposed in our report only apply to these interfaces where natural language is involved.

A number of problems of interfacing arise directly from the circumstances in which language is used. A separate review of these problems might usefully be made. One thinks of the shift towards the spoken word away from things which were previously done on paper, for example the widespread use of the telephone and the development of teleconferencing, whilst on the other hand there is the association of picture and sound, notably with television.

Mixtures of words, sound, written text and pictures are a recent phenomenon known as hypermedia. They are often generated electronically. They react to the use of language and influence its evolution.

To sum up, whilst work on the design of language products *per se* is clearly necessary, it cannot be done in isolation or separated from the impulses it gives to or receives from the world of computer science, notably the communication, storage or processing systems made possible by the new information technologies. All this forms a whole and underscores the legitimacy of the area covered within the Commission by DG XIII.

The market is not a satisfactory regulator of language matters

There would be very little reason for a Community policy if one could count on market forces to control the evolution of the various mother tongues satisfactorily. Why not, you might say, leave it to competition, which would certainly be fierce once it became apparent that language industry products were a profitable business for pioneering enterprises?

Unfortunately very serious surveys conducted by the EC Commission (DG XIII/B) have shown that private initiatives will vary very greatly depending on the frequency with which the various languages are used. There is also a dearth at present of would-be suppliers of language products in private industry.

The Community was quick to recognize this fact. The launch of the LRE(1) programme by DG XIII/B is one result of that realization.

---

(1) LRE : Linguistic Research and Engineering

Japan has also recognized the high stakes to be played for in language technology and is putting a lot of resources into it. Interest in the USA has revived, following a slowdown after the 1970 report "*The present status of Automatic Translation of Languages*" by Y. Bar-Hillel, thanks to the efforts of the big corporations which, finding themselves deprived of adequate data and communications, decided to pool their resources and make their research findings generally available.

The reluctance of private industry to venture into pioneering activity stems from the considerable amount of basic linguistics work which has to be done before the results can be commercially exploited. It also stems from the separation which exists virtually worldwide between linguistics researchers and applications engineers.

If the problem of the language infrastructure of the various countries were left solely to market forces and the distortions which would certainly be caused by Japanese and US policies, it is likely that only a very limited number of the natural languages "alive" today in the world would survive as vehicles for scientific and technical knowledge, as aids to trade and as factors in the propagation of culture. There is no escaping the fact, however, that this is an area in which the scale of investment will determine the scale of the effects. The so-called "minor" languages, spoken by fewer than 40 to 50 million people, will only be able to modernize themselves if conscious measures are undertaken to that end.

Laying the foundations of a language infrastructure

The conclusion of these introductory remarks is that, in addition to the various non-material infrastructures (regulatory, judicial, organizational, monetary, etc.) and material infrastructures (transport, telecommunications, etc.) which an economic area must necessarily have, there is also a language infrastructure.

Revolutionary forces - the only word to describe them - are at work to modify this language infrastructure either for better or for worse depending on the political response which this situation calls forth. The area concerned, as we have seen in the previous section, is that of the "language of technology", but this area determines the ability of the private and public sectors to be innovative and competitive. It is thus vital to the economic and social health of the Community.

To move on from this concept to action, to adapt this language infrastructure by means of a suitable policy, we have to overcome a major psychological hurdle: we are accustomed to the idea that our mother tongue is a gift of providence requiring no particular measures to keep it healthy. To some degree, alarmed by change around in all we saw, we regarded the independent existence of language as something sacred which had to be protected against all artifice. Except in the domain of the "literary language", our analysis has led us to conclude precisely the opposite. Due note needs to be taken of this. Examples of achievements in the area of infrastructure are given in the remainder of this report.

An opportunity for Europe

The threat which hangs over the mother tongues of the Community as a result of the changes which are a feature of our time is in effect an opportunity for Europe. Without the resources of new information technologies, the language barrier might well have been a lasting and virtually insurmountable obstacle to communication amongst Europeans. But if we can manage to find a suitable answer to the challenges raised by the technological transformation of language, we shall have accomplished a decisive step towards a common capacity for communicating amongst different-language communities, whilst at the same time preserving their identities.

So the question is no longer whether or not we should act. Clearly we must. Having got beyond the question of principle, we need to know how we should act: what structures should we use and what should it cost? We attempt to answer these questions in our report, which cannot be more than a first step towards implementation of an ongoing strategy.

The need for a policy

Briefly, there are three reasons in favour of large-scale political action in the area of language use. Those reasons are economic, social and cultural.

1. *The transformation of language tools affects the economy:*

- directly, by the development of industrial and commercial activities related to the language industries. There will be no limit to the spread of language products because they are relevant in all areas of human endeavour. This massive spread hints at big business for the future, though quite when that will be is not yet known. After the optimism of the sixties and the pessimism of the seventies, after a slow start, all the indications are that we are entering upon a phase when specific and very precise applications, particularly monolingual applications, will become a reality.

- indirectly, in that natural language, whether written or spoken, is the main instrument of business management. In business and industry, government departments or any other forms of corporate life, language is the main vehicle for advances in productivity.

  It is becoming indispensable for small and medium-sized undertakings and government departments to have language handling systems, usually monolingual, which offer a performance similar to that already enjoyed by the big multinationals.

2. *Future social patterns will be profoundly affected*

Without suitable policies, the impact of technology will be unequally absorbed across the economic board, depending on the relative dominance of the various natural languages, some of which will have every opportunity to develop, whilst others will be unable to benefit from the technological revolution. If that happens we shall certainly end up with a two-tier language situation.

It is hard to imagine a worse state of heterogeneity when we know how strongly people feel about wanting to receive instructions in their mother tongue. And despite the increasing focus of education on language-learning, one cannot imagine a whole population having to use a foreign language in order to gain access to information which it will increasingly need and which will be distributed by remote methods and on CD-ROM.

3. *The cultural dimension*

A mother tongue is not just a tool of communication. It is above all a vehicle for the elaboration of thought and intellectual and cultural expression. Though understood only imprecisely by most individuals, this notion is strongly held by peoples. We cannot build tomorrow's Europe if one of the main means of expressing its cultural diversity is allowed to wither and die.

I. THE FOUNDATIONS OF A COMMUNITY STRATEGY

1.1 The broad lines

The Study Group responsible for this report began by agreeing on the following three general objectives as a guide on which language infrastructure strategy should be based.

- to improve competitiveness and efficiency in scientific, economic and administrative organizations by the development and use of integrated language and technology tools to aid communications amongst their employees and with clients.

- to enhance social and cultural exchanges amongst individuals and peoples in Europe and the world by the development and use of integrated language tools and technologies.

- to protect and enhance the diversity of cultural resources by the development and use of integrated language tools and technologies appropriate to all the languages concerned.

1.2 A Community policy is justified and consistent with the subsidiarity principle

The Study Group then considered whether action at Community level was warranted. *It concluded that a first prerequisite for success was that each Member State would have to be committed to doing technological research and development on the mother tongues used within its borders.* Application of the subsidiarity principle does not, however, mean that nothing needs to be done at Community level. The Community's role should be to lay the foundations of an infrastructure which in many respects will be essential to the future of the European Union. The Commission must put forward measures to stimulate and coordinate initiatives and harmonize activities by means of standards.

The Commission must, then, act to give this issue a higher profile with leading economic and political figures and enlist their help. We cannot ignore a psychological difficulty, in that the work to be done falls within the realm of the non-material, software, services, and thus lacks the visible impact of, say, prestige buildings, transport or telecommunications systems or space probes. Nevertheless it is vital for every Member State government to view this area as a priority and fund it accordingly.

Given that the question currently has a low profile it will, admittedly, only be an electoral issue in regions where the language issue is critical. But the temptation to take refuge in one's linguistic identity may intensify in all Member States if the economic climate is poor, for example when there is persistent unemployment or increasing discontent in some sectors (farmers). That will then affect the way people vote. The protection of mother tongues is indubitably a matter for a European policy.

1.3 Partners in a Community policy

Commission measures are only justified if they have the backing of the people they are designed to benefit. Thus, before addressing the question of ways and means, the Study Group looked at who the Commission's partners would be. Who should it deal with in implementing a policy aimed at developing a satisfactory European language infrastructure?

The problem previously encountered at the time of the evaluation of EUROTRA in 1989 was confirmed. Spontaneous applications are few in number. 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 invest in commercial products, and users are not creating enough demand for them.

The problem also needs to be highlighted by a general awareness of how critical it is. But this awareness is not keen enough in the mass media or, as a result, in the public mind.

1.4 To encourage progress

As in all things, the various phases of innovation and their connecting areas must be addressed: research upstream of applications, applications research, development, demonstrations, stimulation of the market, increasing user awareness, training of specialist personnel at all levels.

But conventional methods, namely the catalytic effect of precompetitive cooperative research, are not enough. In order to succeed we need to ensure that areas which are currently areas of research become true scientific disciplines (theoretical and computational linguistics, artificial intelligence, phonetics, etc.) and we must get them to interact on an organized basis and then move on well beyond research.

In recent years we have begun to see some spontaneous movement between specific areas and a more interactive approach, thanks to EUROTRA, but specialists from the various sectors of research do not get together often enough. Sometimes it seems they almost deliberately ignore each other in order to protect their specific subjects or intellectual property rights to their work. Links between government-funded and academic research and industry research are particularly weak in this area and there is no tradition of cooperation here. Barring all-too rare exceptions, there is no systematic training or recognized qualification for language engineers, specialists in the interfaces between basic research and applications. But since necessity is the mother of invention, we are beginning now to see some generalists who can make these transfers because they understand the area of overlap between all the scientific disciplines involved and the emerging needs of the market. But such people are rare, and they are self-taught, starting from one speciality and acquiring their additional knowledge by their own efforts. Measures at further education level are thus called for here.

1.5 Structures

The structures will have to allow for the immaturity of the area to be opened up. The situation is reminiscent of the early days of computers in the late fifties. It is acknowledged that by having failed to mobilize intra-Community cooperation in time, the Community is now suffering from a lack of industrial muscle in this area. As regards a language infrastructure we are far from being able to define objectives *a priori* and to know and readily mobilize our partners. This explains and justifies the emergence of conscious policies in the USA and Japan.

Quite apart from implementing scientific and technical and research and development programmes, we thus have a great deal to do in the area of training specialists and generalists on the one hand, and, on the other hand in motivating industry, not to mention persuading users to take an active interest in the preparation of programmes.

The immaturity of the field means that projects are guaranteed to be long-term. We have to take a long-term view and make sure that structures can withstand the short- and medium-term threats from budget constraints and political choices which are bound to arise. This question was reviewed in a preparatory study, the conclusions of which are set out in the Coltof report. Our report proposes a number of possible structures for coordination at Community level. These structures entail a hierarchy of delegated responsibilities to be attributed at each level of action (cf. Chapter V).

1.6 Intellectual and industrial property

The immaturity of the subject is again reflected in the difficulty of resolving the questions of intellectual and industrial property arising in connection with the European language infrastructure. How far should one adhere to the prime view that anything to do with language is in the public domain and must thus be free of all complications over ownership?

How far can one guarantee, by means of individual royalties, a proper return for industrialists who have funded pioneering work and taken the risks that go with that? The Study Group sets out a number of guidelines, with numerous provisos and caveats.

1.7 Funding

The level of funding needed, as outlined in the relevant part of the report, may seem to some to be high. It is reasonable, however, based on what we know of the scale of the efforts which will need to be expended if we are to avoid operating below the critical threshold.

We must not fall into the trap of comparison with the industrial age when anything material was expensive and the physical work of transcribing the work of the mind was relatively cheap. The age of information and communication upon which we are now entering reverses these factors. But we still have difficulty in taking on board the fact that software research demands a lot of human and computer resources and is thus expensive. Projects to develop language resources and basic tools also need to be financed.

1.8 The need to show prompt results

The Study Group was constantly aware of the need to demonstrate progress at all times which, in the event, means that research findings have to be translated as rapidly as possible into useful or marketable products.

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 15 to 20 years later. This is why the recommendations of the Oakley report were adopted for reintroduction into the overall strategy for improving the SYSTRAN system of machine translation. It is also the reason why we propose a structure which will guarantee that partners from industry are involved right from the start of the new programme. It is also the reason why particular emphasis is placed on monolingual applications.

Accompanying measures are also proposed to launch measures which will give concrete form to the concept of a language infrastructure (multilingual communication systems, measures to encourage the wider use of SYSTRAN, etc.).

1.9 Liaison with other programmes

Promotion of a language infrastructure policy for the EEC must form a coherent whole with the efforts pursued by the Commission under framework programmes on research and development and other programmes such as those in the area of education.

The appropriate departments of DG XIII should confer on how best to divide up the requisite work between precompetitive projects carried out under ESPRIT and language engineering, in such a way as to make best use of the common areas of interest which must exist in the various programmes.

But the use of mother tongues is also linked to their presence in stores of information held in databases and banks, on CD-ROM and other disks, in professional and practical information distributed by remote networks and incorporated in software. The structures must therefore provide for liaison with the Commission's specialist programmes, notably IMPACT.

For the rest, efforts aimed at optimizing the results as a whole must concentrate primarily on dividing up the tasks properly amongst the national levels which will be the main bases of the work done. Each language must draw on the progress achieved by its neighbours to make its own technical improvements and prepare itself for machine translation systems. Any large-scale cooperative ventures, under the EUREKA programme for example, must be seen as parts of an overall strategy.

II. THE PARTNERS

In this area which is a highly technical and specific one, the role of the Commission is not to attempt to solve the problems itself but to get those problems identified and solved by the parties to whom they are relevant. The parties to whom the Study Group would point most especially are the users. It is those who benefit from advances in language engineering, who use the language infrastructure and are the consumers of the products offered by the language industries who ought primarily to be involved in helping to plan strategy. This first observation has repercussions for cooperation and consultation structures.

More generally, encouragement must be given to all those able to contribute to the success of the venture, namely national governments, specialist researchers and engineers, industrial manufacturers of language products and, as indicated earlier, users.

2.1 National policies and programmes

In contrast to other areas such as energy, information technology or the environment, the EEC Member State governments have so far done little in the linguistic field towards developing a specialized language industry. Apart from its own initiatives in the form of the LRE and ESPRIT programmes, the Commission thus faces either a void or a few scattered initiatives. The main efforts achieved at national level are limited to a few scientific research ventures prompted by the EUROTRA programme and specialist research projects in phonetics and artificial intelligence. More recently there have been some initiatives aimed at the further development and exploitation of research findings.

One might mention here the interest shown in Germany by the Research and Technology Ministry in the "Verbmobil" project and the involvement of some governments in EUREKA projects. But there cannot really be said to be any strategies or special structures in place at Member State level.

It would be a very good thing to remedy this situation so that Commission officials could enlist the support of national centres - offices, institutes or agencies which would centralize and coordinate all ventures beneficial to the language infrastructure and would have resources allowing them to operate in a way similar to Community bodies. Community initiatives would then merely complement what was done at national level.

The structure in place would make provision for a network of these national centres to operate at Community level if the governments so wished.

This response in the form of national measures is necessary if the subsidiarity principle is to operate satisfactorily. The supposition is that adequate policies will stimulate and coordinate all the public, semi-public and private institutions specializing in this field in each Member State and that an appropriate industrial policy will gradually emerge.

Pending the appearance of concrete national responses to the problems raised by this report the Commission, given the urgency of the matter, should be empowered to proceed and give a lead with initiatives of its own.

2.2 The research institutions

As in every technical field, there has to be research upstream of the applications concerned. This is called fundamental or basic research. Particular attention has to be paid to this part of the initial segment of the innovatory process. European research compares well both quantitatively and qualitatively with that of the USA or Japan. The complication arises from the exceptional and inevitable diversity of the research work done, because successful applications will require convergence and coordination of findings in a large number of specialist areas - theoretical linguistics, computational linguistics, phonetics, shape recognition, cognitive psychology, etc. - which will also have to be related to research in computer science proper (power and speed of memories and handling, parallel structures, signal analysis, etc.).

Researchers are well used to intra-European cooperation. They have cooperated on computer and artificial intelligence topics under programmes such as ESPRIT and, on computational linguistics, under EUROTRA. But there are very few firm bridges between disciplines, and the need for cooperation by two or more sectors is seldom met. Since national structures are not well designed to provide an interface between basic research and applications, this is the weak link in the overall innovatory process. It is a shortcoming which needs to be remedied quickly.

2.3 The language industries

The objective of these industries will be to devise, develop and distribute five categories of language engineering products.

2.3.1 Developmental tools for intelligent text handling systems. Automatic correction of spelling mistakes caused by semantic and syntax error. Elaborators of "controlled" texts (1) (vocabulary and grammar defined a priori). Hypertext elaborator systems with controls for mixing text, graphics and images. Tools for optical scanning and conversion to other print modes and formats, etc.

2.3.2 Tools for the natural-language operation of command systems for databases and banks, conversation with interactive electronic systems. Filing tools (2).

2.3.4 Machine translation systems (SYSTRAN/LOGOS/METAL/SMART/TOVNA). If the European languages are to flourish it is vital that information published in those languages should be readily translatable into the reader's mother tongue. If that facility is not available, researchers and intellectuals wishing to be read worldwide will prefer to express themselves directly in a language which is not their own but will ensure them a worldwide audience.

2.3.4 Tools for automatic data search. Automatic indexing; automated keyword search. Tools for syntactic and semantic text analysis. Tools for automatic relay of information to selected recipients. Automatic watch on e.g. new scientific, technical, financial and economic material.

---

(1) of the SMART type used in 1989 at 150 sites but in English only (cf. Annex II)

(2) for references to systems in operation see Annex II

2.3.5 Tools for conversion between the different forms of natural language (written to spoken). Optical scanners coupled to a voice synthesizer. Tools for recognizing speech and converting it into the written word. Command tools for voice-operated vehicles, robots, medical instruments, etc. Tools for speech analysis. Conversion of written text into Braille. Aids for persons with a sensory or motor handicap, etc.

It will be noted that of the 5 categories of products distinguished by OVUM Ltd (1), 4 relate to monolingual applications and only one to machine translation.

The term "language industries" is new and still controversial. It is used in this report for the reasons listed in the introduction (cf. Section 0.4). The field they cover is defined in Section 0.5. These industries produce and market software specifically for the computer processing of natural languages and their applications. They are quite separate from publishers who produce and distribute books, films for the cinema and television and newspapers. They should not be confused with data bank servers, or with suppliers of software and computer services. But they are part of the same branch of information services. For this reason language industries are usually found as subsidiaries of companies which make computers or sell computer software, or editors, or data servers.

---

(1) OVUM Ltd

Europe's computer manufacturers and distributors are currently experiencing grave economic difficulties. Those in charge are thus reluctant to take the pioneering risks entailed in the language industries and commit themselves to long-term investment.

It is because of these medium-term structural difficulties that the big industrial groupings will not themselves address the need for a good European language infrastructure. When these companies pay detailed attention to the applications of language engineering they do so in order to solve their own internal document handling problems. They are having to do this more and more. But it goes no further, and there are very few who seek to extend and market their results outside the company. There is a danger in that, because this superior efficiency in document handling will widen the gap between them and the SMEs/SMIs which cannot develop comparable services. That superiority further intensifies the "effects of scale".

Independently of the large industrial structures there are many language products which can be developed and marketed by small or medium-sized companies. The potential pioneers here may be members of research teams or former employees of the big firms. Special measures are suggested in this report to encourage this kind of initiative.

Specialist telecoms and network operators can be involved in developing the language industries. They must be encouraged to tender for the supply of specific language services. They are very well placed to understand what a good European language infrastructure should be and make others understand it too (cf. Section 3.5).

2.4 Users

Remember that *the overall strategy proposed for this area is "user-driven".*

Those who stand to gain from the proposed strategy are the 100 million people in the Community whose work consists of processing information, notably the 18 to 20 million employees who regularly use word processors. But in relation to defining objectives, this group is more or less totally passive because it has no conception of the advances which are possible. This requires measures to create awareness and provide training.

The Commission has conducted a variety of studies of this market and in recent years has assembled a large amount of documentation. This study of requirements in conjunction with a study of the possibilities provided by technical progress needs to be pursued further. Ideally a panel of experts should report to the Commission every two or three years on the state of the art in Europe and the other advanced countries and the degree to which solutions have gained acceptance with users.

The difficulty in these analyses is that users are so widely spread in terms of geography and sector of interest. The obstacle can be overcome by a judicious subdivision into categories: manufacturers and servers of data banks, professional translators, librarians, corporate and government documentalists, writers of technical manuals and software suppliers can all make their specific requirements known in detail and provide pointers to the direction which applied research and the design of pilot or demonstration projects should follow.

There will certainly be a psychological dimension to the relationship with users. Remember the initial resistance of many secretarial staff to word processors which are now regarded as an indispensable modern tool. There should be liaison with a group representing professional translators. Instead of instinctively wanting to defend their traditional way of doing things, translators will quickly realize that language technology enhances their profession and extends its applications.

This relationship with the market will of course be gradually transferred to the language industry manufacturers once they offer enough innovatory potential, though this must not mean that the wishes of the end-users are no longer heeded.

2.5 Educators

Any new technical process makes it necessary for staff to be given specialist training. National education systems have done relatively little to train the trainers, the specialists and specific users in the area concerned. Except for a few universities in Europe there is nowhere which provides a training or qualification for "language engineers and technicians", hence the poor degree of interfacing between theoretical research and applications.

Managerial staff in industry and government have never, in the course of their studies, been made aware of the concepts of language industry or language engineering. Their minds make no connection between language and the idea of economic and social infrastructure. Hence the difficulty of getting across to them the importance and the urgency of the problem and the enormous opportunities which exist today for maintaining the linguistic diversity of Europe without allowing it to impede communication.

III. OBJECTIVES AND RESOURCES

The Commission's role can only be to encourage, stimulate and coordinate, as indeed can that of the governments of the Member states. It must also regulate wherever all partners deem it necessary to have a policy of harmonization enforced by standards.

3.1 Creation of awareness and the fostering of a readiness to act

We would reiterate here what we said in a number of earlier sections (1).

One of the Commission's main priorities should be, together with suitably chosen experts, to raise the level of awareness on the part of those in positions of economic and political responsibility in the Member States and create a willingness to act at all essential levels of decision-making. There is no doubt that at the current stage of information awareness it is not so easy to get across the idea that we need, using appropriate means, to resolve the problems entailed in a European language infrastructure, in creating and developing language engineering and promoting a language industry.

The reluctance of some circles to concede this necessity is reminiscent of government and industry to commit themselves, 30 years ago, to the promotion of computer methods or the biotechnologies, whilst the Americans and Japanese were busy launching the programmes which were to establish the superiority they enjoy today.

The subsidiarity principle requires a commitment from Member States. Each government of the European Community has a duty to defend the mother tongue of its citizens.

---

(1) cf. Sections 0.7, 0.8, 0.9, 0.10,1.2, 1.3 and 2.1.

The Commission should only need to provide the additional measures needed to coordinate initiatives, harmonize analysis and processing of the results and cross thresholds whenever effects of scale played a decisive part.

Exceptional measures might be taken to help less favoured regions which were unable to raise the resources needed to defend their language on their own.

The main issue today is to create a readiness to act at the various national levels and put in place in each Member State the appropriate structures and resources.

There are, however, several factors which argue in favour of large-scale action on the part of the Commission itself:

- the need to solve the problem of the language barrier following the Maastricht agreements which endorse progress towards a European Union.

- the need to find an appropriate response to competition from outside the EC, notably from the USA and Japan.

- the need to overcome the inertia which is to be feared at the various national levels.

For this reason it would seem necessary for the Commission to give a conscious lead in a programme well equipped to meet the challenge and move the whole body of solutions along, scaling down its part in the overall effort later once it is possible for control to be taken over at national level.

3.2 The range of possible actions

As we have already said, the area to be covered is the entire chain of innovation from basic research upstream of applied research to the promotion of specialized industrial activities. Each stage needs help in itself, but even more help needs to be given to linking up the successive innovative stages one with the other.

Industry must also be encouraged to apply to help develop prototypes for monolingual and multilingual applications.

There are models. The "concerted initiatives" launched between 1973 and 1981 yielded excellent results in a number of fields. The Multilingual Action Plan (MLAP) evaluated in the Oakley report dates from this period (1977). The major achievement is the introduction of SYSTRAN and EURODICAUTOM by the Commission. This undertaking, which provided a wealth of experience on the difficulties of setting up a CAT(1) system within an organization, is likewise analysed in detail in the Oakley report. The benefits of an operation such as EURODICAUTOM are also commented on in depth.

But the decade of the eighties was the most instructive. The launch of the series of framework programmes, the first fruits of which were yielded by ESPRIT, opened the way.

The EUROTRA programme has put down the foundations of fruitful cooperation amongst researchers in computer linguistics in the 12 countries of the Community. Interest in ventures relevant to the language industries has also been shown by industry under EUREKA.

---

(1) CAT : computer-assisted translation

As regards the problems of nurturing fledgling ventures by entrepreneurs who are still new to pioneering, excellent results have been obtained by the Commission's task force for SMEs and by DG XIII with the SPRINT and VALUE programmes. Valuable experience has also been gained by NATO in its "Science for Stability" programme which might usefully be taken as a model. *We must, in fact, consciously go beyond the bounds of precompetitive research and launch ambitious pilot projects which will lead to directly useable and, if possible, marketable, products.*

These pilot projects will be the key to the greatest challenge, that of creating a continuous chain of innovation, with no weak links, which starts firmly rooted in established basic research and ends in the distribution of products on the market.

3.3 Stimulation of research upstream of applications

We described in Section 2.2 the encouraging position with regard to research upstream of applications. We must make use of these favourable starting conditions and make further progress in relation to our main world competitors, resolving the problems already mentioned over links between sectors and optimum exploitation of results. *We must gradually change the way researchers think and make them more aware of what the market needs.*

This area of research activity upstream of any profitable activity must be supported because it is the source of future advance. It cannot develop without specific funding.

The EUROTRA programme has played a valuable role in stimulating growth at national level, especially in the smaller countries, and in precipitating cooperative ventures. But its underlying principle means that it is a "directed" research project, and that means corrections.

After consulting a number of research teams and taking account of the structural debate (Coltof Committee), the Study Group would like to see greater freedom for researchers in this area in the sense that several schools of thought can put forward competing solutions in order to open up for the future avenues which have hitherto been little explored. Even so it will still be necessary to coordinate the work done. That can be done by means of an appropriate consultation structure (cf. Section 5.2).

This "upstream" research should cover all the methods of automatic language processing offered by the new technologies. Some of the human and material resources are currently being mobilized for EUROTRA and the LRE programme, and others are included in ESPRIT and other Commission programmes. It is not essential to transfer these into one single programme, but there should be an adequate structure responsible for giving unity and coherence to this body of work.

3.4 Pilot projects and demonstrations to promote applications

Industry must be suitably encouraged to change its thinking so that it stops sitting back and waiting and commits itself to pioneering ventures. The classical formula whereby true cooperation between researchers and development engineers can prosper is to devise objectives which are industrially practicable and can be realized jointly.

This is the role of pilot projects which are generally overseen by a leader from private industry. The operation is conducted on a shared-cost basis and lasts on average for 2 to 4 years.

It defines *a priori* phases which are supervised by the "subcontractor", the "main developer" being the project leader plus an adequate consultation apparatus. These pilot projects are also demonstrations organized with the aim of wooing future users.

The first task of a language infrastructure policy should be to draw up, jointly with partners in industry, a list of pilot projects consistent with objectives which have a good chance of success on the market or projects which meet a direct need such as the need to improve the Commission's internal data processing and communication system.

These pilot projects, it has to be said, are expensive, because they involve both research proper and development plus demonstration work and market research. The human resources required are more than a hundred man-years for the biggest projects. The cost breakdown is hard to define in absolute terms and it is better to opt for flexibility rather than a strict 50/50 division. It may be that the "main developer" wishes to make the results available to a large number of interested parties without making a charge for the intellectual property rights. The converse might also apply if the result hoped for is clearly going to create a profitable market for the industrial manufacturer committed to a given project.

The Study Group believes that calling for tenders will be a fruitful source of pilot projects if one can be sure that appropriate budget funds will be available. If that is the case it is likely that a large number of industrial concerns which have hitherto shown no interest will apply; after that it will be a question of choosing the best ones in accordance with carefully defined rules.

A strategy for pilot projects will have to be accompanied by an assessment of each operation's chances of success. Here we come up against effects of scale. If it is to have the necessary dimension, every project, even those concerned with monolingual applications, will *a priori* have to have a trans-European dimension. It will also be necessary to guard against too much competition on one and the same subject, because spreading funds too thinly would clearly mean that the required objective was not attained and the overall project failed. On the other hand, given the immaturity of the field at the moment, it will be necessary to place trust in pioneers and sometimes in small teams. So difficult choices will often have to be made, and there will sometimes be conflicts of interest. That is why the question of the decision-making structure is so important (cf. Chapter 4).

Special measures will have to be considered in order to allow the smallest countries whose natural languages are spoken by only a small proportion of the population of Europe to be involved in the devising, implementation and exploitation of the pilot projects and demonstrations, the effects of which will be relevant to them later.

3.5 Support for instruments of language infrastructure from accompanying programmes

In the course of the surveys which preceded the work of the Study Group proposals were made by certain representatives of business and industry. These proposals help to indicate what form infrastructure projects should take.

A first example is the availability of SYSTRAN translation on Minitel. The user electronically sends the text for translation to the central computer. After a length of time which will vary according to the difficulty and length of the text he will get back the translation in the language of his choice. He pays for the operation through the KIOSQUE system. The cost will depend largely on the density of traffic. If the translating machine is underemployed or if the vocabulary is new to it, the cost and time will be excessive and will put the client off. If the price set by the supplier is set too low in an attempt to get the market going, it is the supplier who will end up discouraged if the flow of clients is slow in coming. In such a case, financial help during the launch and run-in period may be crucial. Based on this experience, a large number of extensions can be envisaged. For example the provision by telephone networks of a fax translation system. The document would be faxed through a translation system and returned to the caller in the language of his choice. The translation could be a raw output direct from the system or, at a higher cost, could be postedited by a professional translator.

Another example was suggested by a German correspondent for telephone calls and teleconferences. Until such time as electronic systems can provide simultaneous interpreting of conversations, human operators could come on the line and make it possible for e.g. a Dane to talk to a Greek when neither one of them had a language in common.

In the category of monolingual applications there are a variety of possibilities for language infrastructure.

Voice-operation of mechanical prostheses represents a considerable advance for paraplegics, tetraplegics and elderly persons with impaired mobility.

Optical scanning in conjunction with automated Braille writing and composition techniques for hypertext provide new aids for the blind, especially in the preparation of school textbooks. The partially sighted can also be helped by synthesized speech based on optical scanning. The hard of hearing can be helped by speech recognition and its transcription into visual signals. Computer tools are very largely independent of the language used. They can be developed in pursuit of objectives common to more than one Member State.

The same applies to education and training, retraining and refresher courses. Language learning, study guidance, training in document retrieval, etc., can be made easier by the use of specialist software in which language engineering will pay a very considerable part.

One could quote other instances of applications concerned with the safety of manufacturing systems and services.

The above examples are given only to illustrate the idea that a large number of infrastructure instruments may be envisaged. Some would be based on remote methods - communications but also document search, research into the setting up and distribution of data banks, content analysis of important documents, etc., whilst others would be monolingual and used for very specific applications which would nevertheless be on an extremely large scale thanks to the Community dimension.

3.6 Constitution of language resources and basic tools

The applications of language engineering depend not only on the results of scientific and technical research but also on the availability of basic tools, the most important of which is electronic dictionaries. Preparation of these dictionaries is a major undertaking. They may contain up to several hundred thousand words for each language, each word being accompanied by a mass of information relating to all the accepted lexical and semantic meanings of that word. This is why it is most important that dictionaries should be reusable and universally valid when one transfers from one monolingual or multilingual family of applications to another. Thus the SYSTRAN dictionaries, which operate in eight Community languages and have no known equivalent to date, need to be preserved when SYSTRAN is re-engineered and should be reusable in other products apparently being prepared at present under the EUROLANG project.

But we have to realize, and this goes for all the basic tools referred to here, that the universality claimed for these tools will only be possible if standards are set and adhered to (cf. Section 3.6 below).

In terminology it will be necessary to incorporate nomenclature lists reflecting developments in the various fields of science and technology such as nuclear physics, new materials, information technology, biology and medicine, aeronautics and space technology, not to mention new offshoots of law, public service, the arts, etc. Every sector creates a considerable number of neologisms every year which have to be collected, defined, classified and translated. The EURODICAUTOM experiment is a promising one and should be extended and broadened.

Other basic tools are concerned with all the mechanisms of form determination and analysis of natural language, grammar, morphological, syntax and semantic analysers, etc. They can also include the creation of an integrated computer environment for document handling.

There is also a need for assessment tools as defined in Section 3.7 below to be developed as soon as possible, for there is as yet no acknowledged standard permitting an objective assessment of projects and products in this field.

Computer technology is constantly producing basic tools which are being developed all the time in the areas of programming languages, filing or textual analysis methods, use of parallel processing structures, etc. These tools are generated by research but need to be refined if they are to be translated into reliable working aids.

It is this area which throws up the question of intellectual property most sharply. Should these instruments which are needed for all natural language processing operations be made available to users simply at cost? Or should royalties be payable on them? This question will have to be addressed in subsequent Commission studies.

Without wishing to give a final judgment on this question, which warrants further investigation, the Study Group thinks that the most liberal solution would seem to be by far the best. Basic tools would be distributed without charge, which presupposes that the cost of developing them would be funded entirely by the national governments or jointly by the Commission and the Member State concerned.

Such a measure ought not, however, to discourage private initiatives or interfere unduly with rights of ownership acquired in the past, which is why the Study Group is giving a general guideline only.

3.7 Standards

Defining standards in a discipline as fluid as language engineering is guaranteed to be difficult. It is necessary if we are to have easier communication between teams together with standardized products, but it must not make for sterility and premature rigidity in concepts which might still evolve further. So this question needs to be looked at very carefully to find a proper balance. DG XIII has already conducted an analysis of lexical resources. We have to go further. At the most elementary level, word processing, ISO(1) has already drawn up a number of standards.

---

(1) ISO: International Organization for Standardization

Standards must include qualitative and quantitative assessment criteria which enable the strengths and weaknesses of products to be identified. As regards quantity, the Oakley report gives an example in a detailed description of how one can count the errors in a machine-translated text. As regards quality, there are test suites similar to those used in computer science to evaluate a system with regard to a given type of application; in the area of voice-operation tools, the SAM project under ESPRIT has broken new ground by devising and developing a range of methods which can be applied to each of the nine working languages of the Community.

It should be noted that definition of these assessment criteria comes under the heading of research, and only after that research is completed can one look to develop marketable tools.

3.8 Technology watch

Embryo technological processes which seem to be promising but do not yet warrant a sizeable investment must be kept under observation by "technology watch". The aim of this is to be aware at all times of the state of the art in the field concerned so that informed judgments can be made later. By way of example there are the different language models of the kind being developed in a number of EUROTRA teams (CAT2, MIMO2, etc.). The Oakley report also recommends technology watch in the area of multilingual automatic indexing, which is the subject of a number of current experiments and which is not covered at present in the MLAP(1).

---

(1) MLAP: Multilingual Action Plan

3.9 Help for fledgling ventures

In a field as new as this, it would be a good thing to encourage initiatives by individuals or small groups who are attracted to the idea of pioneering work.

Financial and human resources possibly drawn from other Commission sectors, e.g. the SME task force, might be used to nurture fledgling ventures (administrative help in launching them, grants and loans of venture capital, etc.) specifically in this area of language infrastructure.

3.10 User involvement

Particular attention should be given to creating greater user awareness and promoting demonstrations which will encourage users to become involved in the preparation of programmes of work in all categories.

Professional translators should ideally be involved in discussions of policy on machine translation.

The surveys we conducted as part of our work on this report reveal a situation of indecisiveness here. On the one hand the professionals want to be involved or would have a useful contribution to make. On the other hand, some of them are afraid of machines taking over their work and possibly threatening their jobs. Regular and open-minded cooperation should overcome these obstacles and lead to fruitful collaboration (cf. Section 2.4).

3.11 Measures in education

The education system did not foresee the evolution of language engineering and thus offers no specialist training. This slowness in responding to a new challenge is nothing new. The same difficulty arose in the past when computer science, ecology or biotechnology were in their infancy (cf. Section 2.5).

As the Member States will probably be varyingly quick or slow to respond, the Commission will have to take appropriate action to:

- encourage the holding of specialist courses, training seminars, exchange of results, etc.

- procure study grants so that nationals of a country where there are no facilities can study in one where specialist training is available.

- inform and put pressure on the Member State governments in order to elicit the necessary response from their education systems, notably the creation of qualifications for language technicians and language engineers.

IV. ESTIMATED LEVEL OF COMMUNITY FUNDING REQUIRED

The Study Group examined the scale of funding which the Community would have to find in order to meet the strategic objectives set out by the Group for the European language programme.

4.1 Preconditions

To avoid any misunderstanding it should be said to start with that our estimates allow for the following facts:

a) Community measures are only a complement to government or private initiatives at national level. *The bulk of the work required to modernize mother tongues must be done by those who use those languages.*

b) Research projects aimed directly or indirectly at scientific and technical progress in this field under other programmes such as ESPRIT, DELTA, etc. will continue under existing funding arrangements and no funds will be switched from them to the language programme.

c) Some funding, notably for training or fledgling ventures, will be make-up appropriations. They will allow the programme to act as a catalyst or precipitator of action, but on the underlying assumption that the customary sources of funding for these measures can meet the cost if necessary (fledgling ventures) or that human and material resources (training) can be drawn from the bodies responsible for developing such resources.

d) If a small country is unable to find the necessary funding at national level, it would receive additional appropriations by the Commission out of special allocations for measures to help the least favoured regions.

4.2 The scale of requirements

The Study Group then considered the overall scale of funding which would be required bearing in mind that, for a five-year programme, the first two years would be years of preparation rather than actual work. We were anxious that funding should be generous enough to yield results which were significant, that is to say more than merely adequate, in each of the fields involved. But on the other hand resources may be slow to be taken up, because of the dearth of suitably qualified people amongst the body of researchers and engineers, and appropriations might be wasted if teams were put together too quickly and without proper guarantees of their quality. The rhythm at which appropriations are allocated will thus be very irregular during the five-year period in question.

In line with the conclusions of the EUROTRA assessment report (1989) the implementation of Community measures ought not to require excessively complex arrangements with governments at national level to share the funding. The normal practice would be that observed in the framework programmes for precompetitive research and not the mechanisms operated for EUROTRA. Even so there will have to be exceptions. In certain cases a prior agreement will need to be negotiated between the Commission and the national authorities.

It is not possible to apply a simple 50/50 rule to the breakdown of funding between the Commission and its partners. As we said earlier, for reasons connected with royalties, the general interest might be better served if certain measures were funded wholly by the Commission. This question remains unresolved, and the Study Group's only recommendation is that it should be considered on a case-by-case basis.

In a number of areas, funding will be limited to make-up funds well below 50% (measures to encourage training, specifically).

4.3 Proposals for the language programme

The Study Group's proposals are set out in the table below (next 5 years)

ACTIVITY	TOTAL MECU	MAIN BENEFICIARIES	COMMENTS
basic research (cf. 3.3)	120	universities, government institutions and non-profit-making associations	equivalent to the rest of the LRE programme but in a framework of wider ambitions
research and development (cf. 3.4)	200	researchers as above + industry	progressive growth after careful definition of objectives
pilot projects, demonstrations (cf. 3.4)	200	industrial manufacturers of language products and users	will produce prompt results
logical accompanying measures (cf. 3.5)	50	partners in language infrastructure projects (learning about practical applications)	prompt definition of a succession of objectives
design of basic language tools (cf. 3.6)	150	language engineers	100% funding? Continuity to be ensured over the term of the programme
prenormative research, development of standardization tools, development of standards (cf. 3.7)	60	everyone	resources will grow in time as the language industries expand
fledgling ventures (cf. 3.9)	10	pioneering industrialists	probably only at the end of the 5 years
training, distribution of results, technology watch (cf. 3.1, 3.8, 3.10, 3.11)	50	trainers, specialists, users	rapid measures needed to train language engineers

The proposed programme would thus require total funding of some 850 million ECU, covering the nine working languages of the European Community spoken in the 12 Member States and a period of 5 years. The cost of the volume of work entailed would average out at something under 20 million ECU per language and per year, or some 100 man-years for each Member State. Given the importance of the stakes being played for, this relatively low level of funding confirms the need for additional funding on the part of the Member States.

The Study Group appreciates that a different group of experts might have come up with different proposals, such is the dearth of operational models at present which might serve as points of reference. There is, in effect, only the LRE programme and very little information about national programmes. American and Japanese programmes do not need to cater for any linguistic diversity within their own national borders. It is reasonable to suppose that other experts might cost the programme at anything between 10% less and 20% more and that they might allocate the appropriations quite differently. One operating method, if consistent with the Commission's accounting rules, might be to set aside a general reserve of 140 million ECU which could be utilized if appropriate after year three, and to allocate 700 million ECU to the various activities listed in Table 4.3 in the proportions recommended by the Study Group.

V. STRUCTURES

5.1 Time scale and continuity

In its debate on structures the Study Group drew on the conclusions of the ad hoc group chaired by Dr Coltof.

The structures put into place will have to be geared to the length of time it will take to modernize the natural languages. We have to be realistic here: Europe's language question will not be solved quickly. An optimistic estimate of the time scale required is 12 years, a more conservative estimate is 20 years or so. And we have to allow for the possibility that this transition period will be even longer thanks to the various technological innovations which will constantly appear and influence the field. After a necessary running-in period the structure adopted will thus have to take account of this time factor and ensure, as far as possible, that work in hand is not disrupted by any stop and go process. Continuity is advocated on several occasions in this report by the Study Group which would like to see a policy of ongoing progress made up of successive phases, each marked by concrete achievements, rather than an attempt at revolutionary solutions which would come too late to a body of users who are ill prepared for them.

5.2 Guidelines for the choice of structures

The Commission has a number of options. All of them have the following features:

a) The language programme is so important that it warrants a budget line which will single it out unambiguously from the rest of the Community's programmes.

b) As the EUROTRA assessment report already suggested in 1989, the decision-making process should consist of two arms:

- the first is political and the responsibility of the Commission. It will define overall objectives, determine the level of funding needed to attain them, coordinate operations with the Member States, deal with participation in related programmes such as EUREKA and prepare a policy on standards.

- the other, "executive", arm will be a counterpart to the first vis à vis the outside world. It will encourage partners, those carrying out projects of all kinds, issue invitations to tender, sift the replies, choose the successful applicants and monitor the performance of contracts. It will be responsible for subcontracted work and for technology watch. In principle this arm will simply delegate; it will not carry out any technical work itself but will retain the right to deviate from that rule if it becomes necessary to exceed certain limits - e.g. by using a big parallel structure computer - which means expending a level of resources not foreseeable at present.

c) This executive arm should ideally involve representatives of the Member States, researchers, industrialists and users in the decision-making process. Its legal form might be that of a European economic interest group, or an industrial and commercial institute or a government agency. The existence of a clearly identified body would be useful in raising the profile of the programme.

d) Whatever the legal form given to the executive arm, the Study Group thinks it should ideally consist of a very small number of permanent staff. Reinforcements could be secured by short- or medium-term secondments from the pool of experts in the Member States.

These would be secured as and when necessary and would make for great flexibility. Another formula would be to subcontract certain duties to the networks of national centres.

5.3 The search for a suitable structure must not hold things up

The members of the Study Group are well aware of the difficulties to be overcome in devising and adopting a satisfactory operating structure. *It is their formal recommendation that the question of structures should not be allowed to slow down decision-making about the launch of a Community language programme and appropriate funding.*

Soundings carried out in government, research and industrial circles in the various Member States reveal a general willingness to see the current structure of DG XIII/B extended to implement an ambitious language programme. This is what our partners on the ground would like. Views were expressed to the effect that technocratic excesses would have to be avoided (objectives and methods decided by officialdom, uncoordinated directives regarding the choice of specialists or cooperation partners, meddlesome checks, etc.). We gather from our discussions with the Commission that DG XIII/B is well aware of this fear and will be able to allay it satisfactorily.

5.4 How structures should evolve at national level

New and changing structures at Community level require matching structures at national level. We pointed earlier (cf. Sections 2.1, 3.1, 3.2, 3.11) to the need for national centres which would be in charge, or at least informed, of policy in each Member State. It is the Commission's job to increase Member States' awareness here.

The political authorities responsible for national centres would be grouped in a suitable cooperative structure forming part of the political body described in Section 5.2 above.

Researchers spread over the various Member States would have to have representatives in the executive body (cf. Section 5.2) to coordinate their research. Industrialists and users would be more specifically involved in defining the objectives of pilot projects, demonstration projects and accompanying measures decided by the executive body. Provision could also be made for them to be consulted over policy on standards.

5.5 Opening up beyond the Twelve Member States

Structural provisions to be studied and put in place should already at this stage take account of the fact that the Community will be opening up more and more towards other countries, notably Northern and Eastern Europe, which will sooner or later wish to be associated with this kind of work. The Commission would do well to retain overall control and supervision of all initiatives and developments in language policy, though the executive body might increasingly welcome new forms of collaboration.

MEMBERSHIP OF THE STUDY GROUP AND BACKGROUND TO THE GROUP'S REPORT

The Study Group was initially constituted as follows:

Chairman: A. DANZIN (Paris)

MEMBERS: H. COLTOF (Amsterdam)
B. OAKLEY (London)
A. RECOQUE (Paris)
H. SCHNELLE (Bochum)

It was assisted by the relevant officials of DG XIII/B, viz.:

R.F. DE BRUINE
F. MASTRODDI
and J. ROUKENS

The Group was assisted in the drafting of its final conclusions by

J. LAVER (Edinburgh)
C. ROHRER (Stuttgart)
and A. ZAMPOLI (Pisa)

Various documents were made available to the Study Group either at the start or in the course of its deliberations. These were:

- *EUROTRA programme assessment reports (PANNENBORG 1988 - DANZIN 1990)*
- *a number of studies compiled for or by DG XIII/B between 1988 and 1991*
- *the final report on the MLAP (Multilingual Action Plan) compiled by a group of experts chaired by B. OAKLEY in 1991*
- *the report drawn up under the guidance of H. COLTOF as chairman of a group of experts commissioned to consider which structures might best meet the recommendations of the EUROTRA assessment reports.*