[*Translating and the computer 13: The Theory and Practice of Machine Translation – A Marriage of Convenience?* Papers presented at a conference … 28-29 November 1991 (London: Aslib, 1992)]

EUROTRA: AN ASSESSMENT OF THE CURRENT STATE OF THE EC'S MT PROGRAMME

Doug Arnold and *Louisa Sadler*

University of Essex

INTRODUCTION

In this talk we are going to give a view of what is happening, and is likely to happen in the 'Eurotra' programme, and its surrounding initiatives[1]. To do this, and to make sure the paper has some relevance beyond the immediate future, we will also sketch out some of the history of the project, and describe the results that have been achieved. We will also give some pointers to the technical, and other literature on the programme.

In outline, the history is this. The official launch of the Eurotra programme was in November 1982, though a certain amount of work had been done by various groups and individuals before this. The stated aims involved research and development (R&D) aimed at production of a 'pre-industrial prototype for the official European Community languages, covering a restricted text type, and domain' (nowadays, we would probably call this a 'feasibility of concept demonstrator'), but there were important secondary aims to do with the development of infra-structure to enable the development of Computational Linguistics (CL), Natural Language Processing (NLP) and computationally applied language studies throughout the member states. From the outset a key principle was the distribution of expertise across the community, and the work on individual languages and language-pairs was carried out by groups of researchers in some 17 University institutes organized into 'language groups' (one per language) across all member states. Fundamental work on the software, formalism, and common linguistic representations was coordinated by a 'Central Team' of between 10 and 20 researchers, drawn from the participating centres.

Though originally envisaged as being of 44 months duration, various delays and extensions due to the accession of Spain and Portugal, (so that the original seven official EC languages

[1] We deliberately talk about Eurotra as a 'programme' rather than a 'project'. A point we will stress is that it is a misconception to see Eurotra as a single project with a single goal, rather it should be seen as a family of related projects. We have tried to make sure the view is at once informed, accurate, balanced and dispassionate. In fact, this is even harder than one might normally expect. Quite apart from the bitter criticism it has sometimes attracted, in order to be well informed about the programme, one must have been close to it at some time, and few people who have been in contact with the project appear to remain dispassionate. We have been involved in the programme for a number of years. We would like to think this has made us informed, while not affecting our ability to have a balanced judgement. Be this as it may, the perspective, and the views expressed here are personal to the authors, and unofficial (in the sense of not being in any way sanctioned by, or necessarily representing the views of, any funding authorities, including the CEC).

became nine)[2] meant that the project was extended until the end of 1990. It was succeeded in January 1991 by the Transitional Programme for Eurotra, the transition in question being that from a pre-industrial prototype to an operational one. Again, strictly R&D aims are accompanied by more general aims, in this case to do with training and encouraging cooperation between industry and research institutes. The transition phase is unlike the original Eurotra programme in making explicit provision for a number of separate and complementary projects alongside the core R&D activity involving contributions from all member states, and in specifically involving non-University research groups. Thus it is a response to the need for diversification in MT and CL research. This phase will end in December 1992, after which it is to be expected that further work building on the results so far (e.g. production of a usable system), will form part of the CEC's programme of Linguistic Research and Engineering, which is itself one of the action lines in the so-called TELEMATICS programme.

For most of its life, the Eurotra programme has had its own funding regime which is quite different from the Community's R&D programmes such as ESPRIT. At the basis of the programme is a series of bilateral Contracts of Association between the Member States and the Commission, and about half the overall budget is directly contributed by the national funding authorities (the precise proportions differ from country to country: the DTI has contributed 70% in the UK). The same regime operates for the core activity by 'language groups' in the Transitional Programme which involves researchers in all member states, while the CEC provides an additional ECU 6 million for funding 'Shared Cost'[3] research, training and industrial participation. After 1992, it is to be expected that work on MT within Community Programmes will largely be funded under the shared cost scheme.

The total budget for the 'pre-industrial' phase of Eurotra (1985-90) was about ECU 44 million, of which about half was contributed by the EC. The EC contribution to the Transition Programme will be about ECU 10 million, and its contribution to the LRE programme is foreseen as about ECU 22 million. To get these figures in some kind of perspective, the EC contribution to ESPRIT (the EC initiative in information technology) in 1987-92 was about ECU 1,600 million, which was itself about one third of the total EC contribution to R&D in this period.

'PRE-INDUSTRIAL' EUROTRA (1985-1990)

In this section we will look at the main scientific ideas, and describe and evaluate the achievements of the work carried out in Eurotra between 1985 and 1990, when one of the

[2] The nine languages are Danish, Dutch, English, German, Greek, Italian, Portuguese, French and Spanish. Notice that the step from seven to nine languages involved a step from 42 to 72 pairs of languages for translation.

[3] In principle, 'shared cost' means that the CEC contributes 50% of the actual costs of a project, with the participants providing the remaining 50% themselves. However, it is possible for academic research institutions, such as universities, to opt for a scheme whereby the CEC contributes 100% of the extra costs they incur through doing the project (that is, salary of non-permanent staff, but not cost of building maintenance, for example).

aims was to produce a 'pre-industrial' (i.e. working, but not practically or commercially usable) prototype system.

The original aims of Eurotra were twofold:

(1) The production of a working, 'transfer based', pre-industrial translation system for the official EC languages, restricted to one subject domain, and text type, but with an emphasis on extensibility, and quality (rather than speed, e.g.).

(2) The promotion of research and education in NLP and MT, and the spread of expertise in these fields, throughout the EC.

In part, (1) was motivated by the CEC's own translation needs, but clearly the other aim (2) went well beyond this, towards laying the technical and infrastructural foundations of a European 'Language Industry', as we now say. Both aims were important in the original project set-up. For example, the second aim motivated the very decentralized organisation, and the way all languages were afforded equal treatment (rather than, e.g. concentrating on the language pairs where the CEC's translation need was greatest) — obviously, there are some respects in which this made achieving the first goal harder.

In evaluating pre-industrial phase of Eurotra, it is important to bear both goals in mind. We think it is reasonable to claim success with respect to (2) – even quite an extraordinary level of success, but one would have to be more modest with respect to (1), even bearing in mind that the aim was not to produce a working MT system of commercial standard, but only a prototype capable of further development.

Scientifically, the most important basic principles were the following (given certain other assumptions about feasibility, they can be seen to be strongly motivated by (1) and (2)):

- The most abstract representations (i.e. those which are input to transfer), the so-called Interface Structures (ISs), are linguistic in nature. As in many other MT systems of this, and earlier vintage, they are based on ideas of Dependency Grammar, where each construction in a sentence is associated with a single lexical item (the *governor),* to which other elements of the sentence are related (as various kinds of *argument,* or *modifier*). Other properties of constructions (e.g. whether a sentence is declarative or interrogative) are represented as *features.*
- Transfer between ISs should be as *simple* as possible, ideally restricted to replacement of lexical material (preserving, as far as possible, structural and featural information).
- Analysis and synthesis should be *monolingual,* in the sense of being independent of which language is taken as target or source.
- Analysis and synthesis proceed via a sequence of intermediate levels of representation, i.e. they are *stratificational.*

Linguistically, observing these principles poses a number of challenges, in particular, theories determining the properties of the intermediate levels of representation must be developed, and an IS theory serving the goal of simple transfer must be articulated[4]. This

[4] Without going into details, two intermediate levels of monolingual representation have normally been assumed: a level of constituent structure, where units are identified, along with their syntactic categories; and a level of relational structure, where surface grammatical relations such as subject, object are represented. Facilitating simple transfer means that IS representations are intended to be 'canonical' in the sense of abstracting away from details of surface grammar such as word order, presence of 'function words' (auxiliary verbs, some kinds of preposition), differences between active and passive forms, and tense marking. The information conveyed by these surface manifestations is typically represented by means of

must be capable of being understood and applied in comparable ways by a large number of linguists working in separate locations, on different languages. From a development point of view, computationally interpretable lexica, and grammatical descriptions of all the different languages conforming to these theories must be produced. The pre-industrial phase of Eurotra produced a substantial, and in many ways very impressive body of linguistic research, the most noticeable achievement being the multilingual nature of the research and the impetus the project undoubtedly provided for work in CL and NLP on languages for which little computationally oriented descriptive work had been previously done. In particular, the theory of IS that was produced seems to be good basis further work developing increasingly 'interlingual' representational theories. The chief limitation here seems to be the restriction to sentence level phenomena, though a certain amount of work on discourse structure has been undertaken, and this will probably be a major theme in the future. One's chief reservations should be to do with the dictionary work — work on linguistics and software was undertaken in the absence of any adequate theory of the lexicon, and it suffered correspondingly. This is regrettable: it means that despite the amount of lexicographic and terminological work that has been done, it is not clear how much of the implemented dictionaries can be usefully reused. Similarly, though to a lesser degree, with the grammatical descriptions, where it seems clear that many of the analyses will be of interest to anyone developing computational applications in the various languages, but it is unclear how far the implemented grammatical descriptions can be exploited in the future.

As regards issues of software, and formalism, the accepted processing model in the early (pre-1985) days of Eurotra was essentially that inherited from systems such as SUSY and GETA, namely a *Controlled Production System,* involving the successive transformation of structures by means of pattern-action rules, which could be organized into sub-grammars under various ordering regimes (hence 'controlled'). However, by early 1985 it was generally felt that this was too unconstrained a model for effective development work in a highly distributed setting such as Eurotra. It was also rather isolated from what was then clearly emerging as the mainstream of NLP, involving unification based formalisms. This dissatisfaction lead, in the early part of 1985, to the development of the '<<C,A>,T>' or 'CAT' Framework. With a few variations and additions, the basic ideas of this framework persist both in the 'mainstream' Eurotra prototype (often called ETS), as well as the 'sideline' prototypes that were produced exploring alternatives, namely CAT2 (Sharp 1991), MiMo (Arnold and Sadler, 1990), and MiMo2 (van Noord *et al* 1991)[5].

features at IS.

[5] CAT2 is the development of an efficiency oriented redesign of the Autumn 1987 prototype implementation undertaken by the Saarbrücken group. MiMo was developed in Utrecht and Essex from early 1987, to explore ideas about the interpretation of compositionality which did not find favour in the project. Instead, the project adopted an alternative development of the CAT ideas, the so-called Mu2e, 'E-framework', or 'ETS' which became the vehicle for 'mainstream' research and development. Perhaps it is worth stressing that the pre-industrial phase of Eurotra actually produced *three* 'sideline' prototype MT systems, exploring significant variants of the 'mainstream' ideas, in addition to the 'mainstream' system itself, and had a significant input into a third line of research in the 'CLG' framework (Balari *et al* 1990). It underlines the fact that it should not be considered as a single project, but an R&D programme with a number of interconnected strands.

The key ideas in these approaches are:

- *Constructivism:* the formalisms distinguish (at least) two rule types: *constructors*, or 'generator-rules', which describe structures at a single level of representation, and *t-rules*, or 'translation' rules, which actually relate different levels of representation. Each level of representation is explicitly described by 'grammar' (a collection of constructors). These constructors are actually used in the mapping from one level to another (e.g. they are used to 'validate' structures that are output from transfer, guaranteeing, *inter alia,* that only well-formed target IS structures can be produced).
- *Compositional Transfer:* the translation of a complex expression is found by combining together the translations of its parts.
- *One Shot Transfer:* each translation rule takes a piece of source IS structure, and yields a piece of target IS structure 'directly' or 'in one shot'. This should be contrasted with earlier 'production system' designs, which involve successive transformation of a source structure by a series of rules, each applying to the output of its predecessor.
- *Homogeneity:* essentially the same formal apparatus is used in Analysis, Transfer and Synthesis. That is, the same sort of rules are used to relating levels in analysis and synthesis, as are used in transfer.
- *Unification:* the fundamental operations of the formalisms involve ideas of *subsumption* and *unification*[6].

In outline, one can see the translation process as follows (we will describe the process as it occurs in the mainstream prototype, ETS). The input sentence is analyzed morphologically and syntactically, by parsing, using the constructors for the surface level of representation. This will give one or more tree structures, whose nodes are decorated with features of various kinds. Each of these trees is translated, via a number of further levels of representation, including source and target ISs, into a surface structure of the target language, from which a target string can be obtained by spelling out the morphological features.

Actual translation between levels involves a number of steps. To begin with, a translation rule is chosen, which matches the source structure in some way (e.g. the rule may be restricted to applying to sentences rather than NPs). Such a rule might be as in (1), simplifying somewhat. In fact, this is a particularly simple rule for the sake of exemplification, so simple that the rule writer would not normally bother to write it — its effects would be obtained by a default translation rule, of which (1) would be just one special case.

```
(1)   A:{cat=s} [ B:gov, C:argl, D:arg2 ]
              =>
      A <B,C,D>
```

The left-hand-side of this rule describes a source tree structure representing a sentence ('cat=s') which has a governor, and two arguments (something like *John kissed Mary* or *The Council of Ministers adopted the proposed Council Decision,* perhaps). The right-hand-side simply says that the translations of the other nodes should become daughters of the translation of the 'cat=s' node. Actually translating the subtrees here involves recursive application of the whole translation process. When all the nodes have been translated, what one has is a

[6] Unification is an operation that combines information from two objects (e.g. representations or descriptions), providing it is not contradictory. To say that A subsumes B is to say that B contains all the information in A, and perhaps something else (non-contradictory) in addition.

collection of candidate target structures. A number of things happen to each of these structures.

First, there will typically be generalizations about the relation between source and target structures (such as the fact that source and target ISs have the same time-reference values). Stating these in every individual t-rule would be redundant (and produce unnecessarily complex rules). Thus, these generalizations can be factored out, and stated in a special kind of 'feature' translation rule (so called because they typically relate features in the source structure to features in the target structure). Variations on this kind of rule deal with cases where a particular correspondence between source and target features is possible, obligatory, or prohibited (in which case the target structure should be eliminated).

Next, the structure is validated by the target constructors. Among other things, these can propagate features around the structure (ensuring, e.g. that verbs and subjects agree, according to the rules of the target grammar), re-order nodes (so that target word order constraints are satisfied), insert fresh nodes (e.g. for function words), and act as filters eliminating certain kinds of structure.

It should be obvious that this is amounts to a linguistically very powerful, expressive, and flexible formalism. In particular:

(i) The compositionality and 'one-shot' requirements make the translation process relatively orderly, systematic, and easy to understand (e.g. in comparison with transformational systems). For example, one is not faced with the problem of tracing the way a representation is transformed by a sequence of pattern-action rules, and the formalism is 'declarative' to an interesting extent (so that the rule writer need not be bothered about details to do with the order in which things happen),

(ii) Similarly, constructivism means that only well-formed structures can be built at any level. One effect of this is to keep things understandable, for example, the job of a synthesis writer is greatly eased by knowing that, in principle, the set of possible inputs to synthesis is defined (by the IS constructors), independent of any particular source language or transfer component. However, this is also the basis of some interesting linguistic strategies, whereby the t-rules can be left rather general, and the target Generator relied upon not just to choose the correct output ('filtering' wrong translations), but 'filling in' information that is target language specific.

(iii) The formalism allows one to achieve a reasonable degree of modularity in one's descriptions. That is, not only is it possible to give a good account of individual linguistic phenomena, it is also in many cases possible that the rules devised for the individual phenomena interact correctly when the phenomena occur together (to take a rather simple example, it is not enough to be able to deal with certain tenses, such as subjunctives, and problems of translating the English adverb *just* into French verb *venir (de),* and embedding of sentences within one another, one must also be able to deal with examples like *If they were to have just left....,* without inventing special additional rules. This is a serious problem, which is often neglected in MT.

(iv) Of less immediate importance from a development point of view, but of more scientific interest, is the fact that this formalism, and its variants, allow one to address a number of important formal questions in a contentful and productive way. For example, issues to do with the correct balance of work between the descriptions of monolingual and translational components, and how to avoid redundancy between them; how powerful does the transfer notation have to be (i.e. how precisely should 'compositionality' be defined)? what is the optimal degree of procedural control for the user of a translation formalism? how can one achieve the right interaction between descriptions of general, regular, or 'default' cases, and exceptions in translation? how far can a transfer notation be 'reversible' (e.g. can one set of rules to be written which are adequate both for English-French, and French-English translation)?

However, there are a number of serious weaknesses, which should be noted. First, the application of this sort of formalism to inflectional morphology is often problematic, and as already noted, it does not have a sufficiently developed view of the lexicon, and lexical processes. Second, though it is in many ways close to what has emerged as the mainstream of Computational Linguistic formalisms, it is not sufficiently close for descriptive insights and results obtained there to be directly incorporated or applicable, in every case. This is especially the case with respect to analysis and synthesis, where the stratificational model is far from ideal. Third, the sheer power and flexibility of the ETS formalism especially means that it is inherently very complex, which makes for practical problems of slowness, so that testing descriptions can be time consuming and frustrating.

In evaluating the achievements of the pre-industrial phase of Eurotra, it is important to repeat that the intended goal was never the production of a usable MT system, comparable to commercially available products. The aims were in part purely scientific, and educational (to be measured with respect to knowledge acquired and disseminated), partly strategic (to be measured with respect to infrastructure developed, and the general state of the 'Language Industries' in Europe), and partly developmental (to be measured with respect to a working prototype, which was to demonstrate feasibility and be the basis for further development).

Taking these points in reverse order, a working prototype was certainly produced – in fact, as noted several distinct working prototypes were produced, each exploring significantly different variations of a basically common approach, and all have been demonstrated in various places. With the possible exception of CAT2 (which is by far the simplest, and most limited of them), none of these prototypes bears any comparison with standard industrial systems on a number of points – notably with respect to lexical coverage, speed of operation, and user interfaces (the last of these is not very significant in a prototype system, which is not intended to have actual users). However, in defence of the 'mainstream' prototype one should note that it display a number of important features. First, it is multilingual to an extent that is almost without equal, in terms of the sheer number of monolingual components that are available, and in terms of the number of translation pairs. Second, it is designed in such a way that if it produces anything at all it is normally correct (i.e. the quality is generally high, as originally intended). Third, a relatively wide range of rather complex, theoretically challenging phenomena are dealt with (such as 'Support Verbs'[7], and the treatment of tense), not just in isolation, but in interaction. This is strongly suggestive of extensibility.

Nevertheless, there are some respects in which the developmental results are in some respects disappointing. By contrast, we think the results with respect to infrastructure are extremely impressive. To appreciate this, one would have to appreciate the sheer lack of understanding that existed between linguists, computational linguists and computer scientist from various parts of Europe as recently as the early 1980's. It was then quite often literally the case that computational linguists from different parts of Europe could not communicate face to face on professional matters. In the early days of Eurotra, this problem was so pervasive that a special term was coined to describe it (the 'dtsb' - different training and scientific background - problem). This problem is now greatly diminished, partly as a result

[7] Support Verbs are the semantically empty verbs in '*enter* a suggestion', cf. Dutch 'een overeenkomst *sluiten'* — literally '*close* an agreement'.

of training, some of it 'on the job', and partly as a result of the effort of trying to work together, there is now a large pool of researchers and research groups which have direct experience, not just of successful long distance communication about professional matters, but of successful direct collaboration on single pieces of work. This together with the channels of communication that permit it is a valuable resource for the future. It is also worth pointing out that there are EC countries where there was essentially *no* computational linguistics before Eurotra, and where almost all existing computational linguists have received their training in Eurotra, or closely related projects.

Finally, as regards the scientific aims, one can certainly point to a large, and in places very impressive body of linguistic research, and to the development of transfer formalisms which bear comparison with any competitors on a number of fronts.

In assessing this work, one should perhaps distinguish a view of the general design, from a view of the manner in which that design was executed (of course, the difference is not always obvious). Regarding the general design, we think one can be rather positive. By and large, we are convinced that it was right to design a multilingual, general purpose, transfer based, non-interactive MT system, with the sentence as the largest unit of analysis, and using essentially linguistic (rather than, e.g. common sense) knowledge. In an EC context, multilingualism (as opposed to approaches based on single language pairs, or single directions of translation) seems the right option, especially if the work is to form the basis of work on other kinds of multilingual language processing. Despite its inherent limitations, advances that have been made in interlinguality, we believe that transfer is still the right basic approach, and the work that has been done in pre-industrial Eurotra does not preclude extensions and modifications to utilise other sources of knowledge (e.g. probabilistic or AI techniques), discourse and text level description, user interaction, or other architectures (e.g. 'example based', or negotiation and constraint based approaches), so long as these involve some rule based component (and we think they should).

As regards the way the design was executed, one's assessment must be slightly more mixed. While there are a number of key issues in transfer based MT where work in Eurotra can be seen as exhibiting a clear advance, or where (as with stratification), it is only recent theoretical developments that make the Eurotra approach seem clearly unsatisfactory, nevertheless, we have noted a number of other important issues which were barely addressed. Perhaps the most serious weakness of all, however, was the organisation structure, within which the tension that always exists between research and development work was never properly resolved. It was a common experience that interesting theoretical ideas were discarded because the necessary research could not be reconciled with the demands of the development schedule, while development work had to be done with linguistics, software and formalism which were still at a research, or experimentation stage (almost all implementation in the project was done with software at no more than a 'runable specification' stage in any case). Fortunately, a clearer separation of research and development work is a central feature of the current Transition Programme.

CURRENT STATUS: THE TRANSITION PROGRAMME (1991-1992)

In this, the current phase, the aims of the programme concern mainly the preparation of an operational Eurotra system and the reusability of resources (lexical, terminological and grammatical).

In fact, the interpretation of the first aim is rather broad – creating the conditions under which an industrial prototype with the general characteristics of Eurotra (i.e. general purpose, multilingual, transfer based) can be developed (this transition from a pre-industrial to an operational prototype system accounts for the name of this phase). This involves a good deal of basic research as well as R&D work and the dedication of resources to training (to develop expertise across all member states).

This broadening of aims from those of the pre-industrial phase implements one of the recommendations of the evaluation of Eurotra-1 (Danzin), and underlines the way one should see Eurotra as a *programme* of interrelated research and development work/projects, rather than a single project. As we will see below, it is intended that this broadening process will continue in the future, when other, related aims will be added.

Before describing the various kinds of activity involved in the Transition Programme itself, a word is in order about the 'feasibility and design' studies which were undertaken in 1990, alongside the other activities of the pre-industrial Eurotra programme (the so-called ET-6 and ET-7 projects). As will appear, these various studies relate to the specific aims and objectives of the Transition programme and feed directly into the work in the current phase, as we shall see below.

The ET-7 feasibility study concerns the development of tools and methods for the reuse of lexical resources in computer applications (Heid and McNaught).

The ET-6 studies were intended to assess the strengths and weaknesses of the current prototypes with respect to the state of the art in CL and NLP and propose an improved framework. A number of high level requirements were placed on the formalism redesign, amongst which that the design had to be totally mainstream and extensible as new phenomena and capabilities can be added. The first of these developed specifications for a new formalism (the so-called ET-6 Formalism, ET-6/1, Pulman, ed), the second led to specifications of a user and grammar development environment (ET-6/2). A third (ET-6/3) dealt with issues of low-level text encoding and handling (including some morphological analysis).

As regards the Transition Phase itself, the following four activities being pursued:

(a) A continuation of R&D work within the current framework, especially contrastive research on linguistic topics, aimed both at improving the current prototype, and providing a solid basis for work in the future, beyond the useful life of this prototype;

(b) Implementation of an enhanced development and research system (formalism, development environment, etc.,) along the lines specified in the ET-6 studies;

(c) Research on a number of key topics, where the costs are shared equally between the CEC and industry (or funding authorities in the member states);

(d) Training, carried out mainly in the participating centres.

Before looking at these in more detail, it is worth saying something about the organisational framework for this work. In part, this is similar to the organisational framework of Eurotra-1, in that work under (a) is being carried out by essentially the same research groups that were involved in Eurotra-1, and on the same funding basis (that is, with joint funding from the CEC and the member governments (ECU 8M). There are a number of minor differences, which are not very important here (in part they reflect a difference of emphasis in the work, e.g. the common linguistic theory, and software are now essentially stable, so that many of the functions formerly fulfilled by the 'Central Team' can now be fulfilled by a (somewhat enlarged) Commission team). Similarly, work under (b) is directly funded by the CEC (ECU 2M), as was the work of designing the original software. However, as noted above, the 'Shared Cost' basis of research in (c) is a novel departure in this context, though it is f amiliar in other CEC initiatives such as ESPRIT. The intention is to involve nonacademic, specifically industrial, groups in R&D work, thereby promoting the kind of organisational framework of academic-industrial collaboration that will be required if an operational system that builds on Eurotra R&D is to be achieved (and which is in any case seem as a necessary part of the development of the 'Language Industries' in Europe). The CEC contribution to (c) is around ECU 3M and the remaining portion of the CEC budget (around .5M ECU) is dedicated to training.

Work on (a), improving the current prototype, and extending linguistic research that will be required for an industrial prototype system, can reasonably be seen as a continuation of the work done in the pre-industrial phase. However, from a number of points of view the structure is better. In particular, the prototype implementation, and the linguistic theory as it relates to the development work have been stabilized, and in general both research and development work are more clearly distinguished, and better focused. Research is being carried out on a number of monolingual, and contrastive linguistic topics. In the former case, the aim is ultimately to provide better treatments of particular phenomena in individual languages (for example, work on English, at Essex focuses on the treatment of idioms and modifiers/adjuncts). In the case of contrastive research, the aim is to provide linguistic analyses which lead to representations where the differences between languages is minimized, on the model of time/tense, where the pre-industrial phase saw the development of a representation based on a language independent representation of time-reference, in place of language specific representations of morpho-syntactic tenses. The contrastive topics receiving particular attention include: determination, negation and quantificational phenomena (which all involve notions of scope); compounding; and aspectual phenomena.

Work on (b) is being undertaken by a consortium of European software companies, essentially producing an industrial implementation based on the outcome of the ET-6 contracts.

The feasibility studies are also important in relation to the 'Shared Cost' work under (c), where they provide the reference formalism, processing model, and environment. The original call for proposals here suggested a wide range of possibilities, from research on translation theory, through general Computational Linguistic topics (to do with morphology, interaction of lexicon/dictionary and grammar), work on knowledge representation and terminology, to work exploring applications of subparts of the existing system. In the event, five projects have been funded (a sixth may be added shortly), all of which will take ET-6/1 as their reference model:

- Collocations and the lexicalization of semantic operations. This concerns the extremely important problem of dealing with collocational restrictions (e.g. '*rancid* butter' vs '*sour* milk') in MT.
- Terminology. The objectives of this project include the definition of the internal representation of terminological definitions and their use in analysis and generation, the (semi-) automatic parsing of definitions and the use of the output of such parsing in analysis and generation.
- Knowledge Bases. This project involves an assessment of the feasibility and effectiveness of the (semi-) automatic parsing of dictionary definitions (from the COBUILD dictionary) as a form of knowledge acquisition for ET-6, with wider relevance for other natural language systems.
- Implementation of Probabilistic and Corpus-based methods in Eurotra. This project will investigate the possibility of adding a probabilistic component to the ET-6 architecture.
- The Reusability of Grammars for ET-6. This involves research on the migration of grammars to the new (ET-6) formalism. This work is of practical interest with respect to

the new formalism, but also bears on the general issue of reusability of linguistic resources, which is likely to be increasingly important in the future. This selection of topics for funding shows the importance currently given to lexical issues in current NLP, and to remedying deficiencies in the existing Eurotra dictionaries, and an interest in complementing the linguistic and rule-based approach which was adopted in the pre-industrial phase, by investigating how other kinds of knowledge can be added to a linguistic, rule based system.

FUTURE PROSPECTS

Looking to the future, one can see a number of initiatives that will draw on the results of Eurotra to date. As regards research, much of the work currently being done, including the development of an improved development environment, and the implementation of the 'new' formalism is promising, the latter because it should bring Eurotra work directly into the mainstream of contemporary computational linguistics. From a development perspective, there is a proposal to produce a commercially usable version of the CAT2 prototype. CAT2 is both the simplest, and in practice most efficient of the Eurotra prototypes, and within its theoretical limitations it could be the basis of a very useful commercial product. A more ambitious enterprise is the 'Eurolang' project, aimed at producing a commercial translation system for English, French, German, Italian, and Spanish whose main driving force is the French documentation company SITE. This is a development project seeking to synthesis some of the results of Eurotra with the long standing achievements of the GETA group. The planned start up for this is early 1992, the project has just received the EUREKA label, and seems broadly on schedule.

Beyond this, though no doubt some of the work that builds on Eurotra R&D will find its natural home in ESPRIT (i.e. information technology), it is to be expected that most of the work will fall under the LRE initiative, and we will concentrate on this here.

CEC initiated R&D is organized into a series of framework programmes, each divided into 'Action Lines'. The Third Framework Programme (1991-4) includes, as one of seven Action Lines under the TELEMATICS[8] program, one headed 'LRE' *(Linguistic Research and Engineering,* area 6), which is intended to support work on NLP that has strategic importance, and which can be seen as the natural successor of Eurotra, both in building on its achievements, and having goals which subsume those of Eurotra – specifically, to support development of basic linguistic technology, "...with a view to overcoming limitations and inefficiencies brought about by the use of different languages within the Community" (LRE Call for Proposals) i.e. overcoming the (multilingual) language barrier. The total budget for LRE is about ECU 22 million, with most of the research being organized on a 'shared cost' basis (though the plans also to allow for some research that is wholly CEC funded, via grants, and study and service contracts, where this is appropriate, e.g. as with III below).

Work under LRE is grouped into five main headings. Calls for proposals under three of these headings (I, II, and IV) were published in November 1991, other calls are anticipated in 1992 and 1993.

[8] It may help some readers to know that the TELEMATICS action line is at the same level of organisation as (though smaller than) the action line which relates to Information Technology (ESPRIT).

I Research of General Interest: Here the main emphasis is on research on: ways of increasing the interlinguality of linguistic representations of text/discourse; the use of domain specific knowledge (e.g. terminological, 'real world' specialist, and 'heuristic' knowledge); interfacing NLP and speech technology; and the use of 'advanced computational technologies' (to ensure that NLP R&D keeps pace with progress in computing).

II Common Tools and Resources: Here the aim is to develop 'generic' software tools, grammars, dictionaries, terminological collections, and text corpora, which can be reused for a variety of applications and purposes (and hence are in some way 'theory neutral'). Typical software tools would be integrated testing and development environments, tools for dictionary construction and use, corpus analysis tools, 'workbenches'.

III Linguistic Standards: Here the aim is the definition of commonly agreed data encoding schemes and formats for linguistic resources (e.g. dictionaries, grammars, corpora). This clearly relates to the last point. The kinds of actions foreseen include the setting up of European 'expert groups' to begin working towards the formulation of such standards, experimentation with existing candidate standards, and support for similar national and international initiatives.

IVApplications: The aim here is to support pilot and demonstrator projects in areas such as: MT; automatic document abstracting and indexing; aids for mono- and multilingual document generation, storage and retrieval; HCI; construction of Knowledge Bases from Natural Language Text; and Computer Aided Instruction. Such projects demonstrate the feasibility, and manner of application of work under other headings, measure progress, and provide a way of testing results.

V Supporting Actions: This covers training, initiatives to raise awareness, gather, synthesise, and disseminate information about NLP, with special emphasis placed on examining the economic and social impact of the technology, and legal problems that act as barriers to the emergence of new products and services.

What one sees here is the intention of overcoming the most obvious problems that beset practical NLP: the lack of common resources (e.g. dictionaries, grammars), frameworks, tools, etc, which make it difficult for one project to build on the results of others; the encouragement of research that addresses immediate, apparently solvable problems, and of development work that shows how this, and existing research can be applied; and investment in training and other infrastructural activity. Though one may have reservations about points of detail or emphasis, this seems essentially sound policy.

CONCLUSION

In this paper we have tried to give an overview of the past, present, and future of the 'Eurotra' programme, which, we emphasise should be seen as a family of related projects with a variety of R&D and infrastructural objectives, rather than a single project with a single (developmental) goal. As regards the past, we think it is important to emphasise that while not an unreserved success in all respects, especially as an environment for carrying out research and development work, it achieved far more than its detractors would have one believe, perhaps more than could reasonably have been expected in the circumstances. In particular, though there are important gaps in the R&D that was done, it has provided a sound basis of research both in linguistics and software/formalism, and an excellent infrastructure of expert personnel and research groups. We think that the changes one sees in the current Transition

Phase are mainly beneficial, especially in the relationship that exists between research and development work, and the diversification of efforts into smaller, but still related projects. The work that has been done towards an improved formalism and research and development environment is also promising. Similarly for the future, much depends on the precise interpretation and emphasis that is given to the LRE goals, but there appear to be grounds for optimism that interesting research, and productive development work will be possible.

It is not to belittle the work of other European MT researchers to say that it is in large part because of Eurotra, perhaps *only* because of it, that Europe maintains a lead over Japan and the USA in MT research. Of course, it is too early to say whether this will successfully translate into a similar industrial and commercial strength, but the foundations are there.

LITERATURE

In this section we will give some pointers to the literature addressing the matters discussed above. Full references can be found below.

A good general discussion of the Eurotra project (i.e. Eurotra-1) can be found in Raw *et al*. For independent evaluations of 'pre-industrial' Eurotra see Danzin, A. (1990), and Pannenborg, A.E. (1987). The primary source of detailed discussion of Eurotra linguistics is the *Reference Manual* (the current, and more or less final version is 7.0, earlier versions are of mainly historical interest). More discussion of the linguistics can be found in Copeland *et al*, eds, and in the special issues of the journal *Machine Translation* dedicated to Eurotra (vols 6.2, and 6.3, 1991). A description of the 'new' formalism mentioned in section 3, (the so-called 'ET-6 formalism', can be found in Pulman (ed)). The results of the feasibility study on the development environment can be found in ET-6 (Schütz, ed). Discussion of reusability of lexical resources can be found in Heid and McNaught (eds, 1991).

Copeland *et al* is in fact just the first of a series of volumes in the *Studies in Machine Translation and Natural Language Processing,* which is intended to embody the research results of the pre-industrial phase of Eurotra and the Transition Programme, in so far as they are not reflected in the Reference Manual. Other volumes in this series will consider: The Eurotra Formal Specifications, Assessment of Computational Linguistic Formalisms, Morphology, Interlevel Processing (i.e. mappings between levels within analysis and synthesis), Support Verbs, the Argument Structure of Nouns, and Preference Mechanisms (i.e. methods for choosing between competing analyses).

Our discussion of the LRE programme is based mainly on the *Technical Background Document* of the Programme, and the relevant Council Decision (91/C 218/04). Discussion of the Transition Phase is based on a variety of documents, including Pugh (1991), and the text of the relevant proposal for a Council Decision (Com(89)603 final). A good general guide to the structure of CEC funded research is *EC Research Funding: a guide for applicants.*

REFERENCES

D.J. Arnold and L. Sadler, "The Theoretical Basis of MiMo," *Machine Translation,* vol. 5, pp. 195-222, 1990.

Sergio Balari, Louis Damas, Nelma Moreira, and Giovanni Varile, "CLG(n): Constraint Logic Grammars," *COL1NG-90,* vol. 3, pp. 7-12, Helsinki, 1990.

Commission of the European Communities, *EC Research Funding: a guide for applicants,* DG XII, Brussels, 2nd edition, May 1990.

Commission of the European Communities, "Call for proposals for the Specific Programme of research and technology development in the field of Telematic Systems of General Interest (91/C 218/04)," *Official Journal of the European Communities,* 21 Sept, 1991.

C. Copeland, J. Durand, S. Krauwer, and B. Maegaard, eds., *The Eurotra Linguistic Specifications,* Studies in Machine Translation and Natural Language Processing; 1, Office for Publications of the CEC, Luxembourg, 1991.

A. Danzin, *Eurotra Programme Assessment Report,* DG XIII, CEC, Luxembourg, March 1990.

Eurotra Reference Manual, 7.0, DG XIII, CEC, Luxembourg, 1991.

Ulrich Heid and John McNaught, eds., *Eurotra-7 Study: feasibility and project definitions study on the reusability of lexical and technological resources in computerized applications,* DG XIII, Luxembourg, August 1991.

LRE Programme, "Technical Background Document," *Call For Proposals 1991,* DG XIII, Luxembourg, 1991.

Jörg Schütz, ed., *Feasibility study for a Eurotra-II system: a feasibility study on the software environment with the framework of the Eurotra programme,* DG XIII, Luxembourg, 1991.

A.E. Pannenborg, *Eurotra Assessment Panel Final Report,* DG XIII, CEC, Luxembourg, October 1987.

Jeanette Pugh, "1991-1992: The Eurotra Transition Programme," *Proceedings of a Workshop on Machine Translation, UMIST, Manchester, 2-3 July, 1990,* DTI/Speech and Language Technology Club.

S. G. Pulman, ed., H. Alshawi, D.J. Arnold, R. Backofen, D.M. Carter, J. Lindop, K. Netter, J-I. Tsujii, and H. Uskoreit, *ET6/1: Rule Formalism and Virtual Machine Design Study,* SRI International, Cambridge, 1991.

A. Raw, B. Vandecapelle, and F. Van Eynde, "Eurotra: An Overview," *Interface: Journal of Applied Linguistics,* vol. 3, no. 1, pp. 5-32, Vlaamse Economische Hogeschool, Brussels, 1988.

Randall Sharp, "CAT2 – an Experimental Eurotra Alternative," *Machine Translation (Special Issue on Eurotra (II)),* vol. 6, no. 3, pp. 215-228, 1991.

Valerio Allegranza, Steven Krauwer and Erich Steiner, eds., "Special Issue on Eurotra (I)," *Machine Translation,* vol. 6, no. 2, 1991.

Erich Steiner, ed., "Special Issue on Eurotra (II)," *Machine Translation,* vol. 6, no. 3, 1991.

Gertjan van Noord, Joke Dorrepaal, Pim van der Eijk, Maria Florenza, Herbert Ruessink, and Louis des Tombe, "An Overview of MiMo2," *Machine Translation,* vol. 6, no. 3, pp. 210-214, 1991.

AUTHORS

Doug Arnold and Louisa Sadler, Department of Language and Linguistics, University of Essex, Wivenhoe Park, Colchester, CO4 3SQ, UK