THE LANGUAGE INTERNATIONAL INTERVIEW

Serge Perschke

Head of the EC Eurotra project

by Geoffrey Kingscott

Serge Perschke, head of the Eurotra programme, the European Commission's machine translation research project, was interviewed in his office in Luxembourg by Language International *editor Geoffrey Kingscott.*

GK: *Could we take you through your own career, I know you were at EURATOM but can you tell me your origins and your language background?*

Perschke: Well, I was born in Russia, of a mixed family. My mother is Russian, my father German, so my first language was Russian, and this was the reason, I guess, that I was attracted to the humanities, and in particular to language studies and specifically Slavonic philology.

GK: *Did you go to school and university in Russia, or Germany?*

Perschke: In Germany, and my knowledge of Russian was actually the reason for me getting involved in machine translation, because, you may remember, in the 50s the only language pair of interest particularly to those who were interested in machine translation was Russian into English, and the main prospective customers for such a translation system were the intelligence community in the United States.

GK: *Did you have a university education?*

Perschke: I did Slavonic Studies in the University of Cologne; and at the end of the studies I got an offer from the University of Milan who had a contract with the US Air Force, and at that time — it was the late 50s — very advanced and revolutionary ideas about what machine translation ought to be, or natural language analysis.

GK: *What sort of advanced revolutionary ideas?*

Perschke: It resembled an artificial intelligence approach, with very strong semantic and pragmatic orientation. Well, the term "artificial intelligence" had not been invented yet at that time, and the state of the art of computing was not really suitable to implement that idea.

GK: *I'm trying to work out where you made the jump from a degree in Slavonic Studies. Presumably there was very little element of linguistics in it because people didn't even talk about linguistics at that stage.*

Perschke: It was not computational linguistics but the Slavonic philology tradition. Classics is very strongly linguistics oriented, and I had more linguistic than literary orientation in my studies.

GK: *When did you first meet the computer?*

Perschke: In Milan, this was in 1959. I knew nothing about computers, but in this project of the University of Milan, we had a few what we called then programmers — we call them now computer scientists — and we found out very fast that not knowing computers it was virtually impossible to communicate with such people. You wouldn't understand what they were telling you and they wouldn't understand what you were telling them, so I had to learn very much about computing the hard way; just trying it, this was the only way of doing it in that period.

Anyway, this project ran out around 1963, during the Kennedy Administration. The reasons were not lack of interest either in the United States or in the method of approach itself, but purely political problems, that the United States stopped sponsoring research abroad. In the '50s they had had a huge surplus in foreign currencies which they used for this purpose and the Kennedy Administration stopped this about mid '63.

GK: *Did you go to EURATOM then or was there an interval?*

Perschke: Then, yes. When the Ameri-

can Government stopped financing the project in the States, a good part of the team went to Ispra to finish the work and the system then remained in Ispra. I participated in this final development of the Georgetown system in '63 and '65. I had had a contract with them since about '60-'61 and when this contract finished I took over the system, actually set up a relatively small scale machine translation service. It was addressed to the scientists who were really interested in the progress of Russian technology and physics and who were quite happy getting — sadly — bad English text rather than nothing, and our approach to promoting the service was to translate by computer, first the tables of contents of the incoming journal, and to circulate them in the centre. People were then asking for the translation of this or that article.

GK: *Still by machine translation?*

Perschke: By machine translation. We had no capacities for post-editing. We did a few experiments asking the specialist himself to post-edit it, without knowing any Russian, and a very high percentage got things right.

GK: *So it's very similar to the experience at the Wright Patterson air base, that people don't want revision if they are experts.*

Perschke: If you're an expert and you want to know what it is about — it is mostly enough.

GK: *What happened then?*

Perschke: This service held up for about ten years from 1965 to '75. What happened was that the Georgetown system was programmed in the usual way with machinery which was assembled for a second generation of computers — IBM 70/90 — and already in the mid-'60s a new generation — the 360 series — was coming up which was totally incompatible with the old system.

I started worrying about the future of such a system, that it would cost a lot of effort to re-programme it for a new generation even without changing anything, but having made or participated in making such a system of course we had lots of ideas on how one could do better...

It was a fairly difficult period: (i) because of the famous ALPAC report which had come out in 1966 — it had actually been written in '64 already but the damage was done in '66 — and (ii), there was a profound crisis at EURATOM in that period. The member states could not agree about what Euratom actually ought to do.

The reason for that was straightforward. Euratom was conceived in the first petrol shock of 1956 and the Suez crisis, which made in fact the European states a centre of energetic effort.

The two reasons came together: (i) in the '60s petrol became extremely cheap so there was no real motivation for pushing this sort of research, and (ii) having had relatively short-term planning these reactor types became industrial reactors and EURATOM was considered competition to industrial companies and development ...

GK: *So there was a crisis in EURATOM, combined with the poor morale in machine translation from ALPAC?*

Perschke: In principle the management of the centre agreed that we should do something, and so on, but we didn't really get the means of launching a real project while these things were continuing.

In the rest of the Community the interest in scientific and technical information at Community level was growing and in 1974 or '75, I don't quite remember, the Community here in DG XIII here in Luxembourg launched an Action Plan for scientific and technical information with two main objectives — the first, to create a European network, dedicated, which then became the first operational packet switched network which led them to the X25 standard, which is fairly well known now; and the second purpose was to stimulate the creation of databases.

As a sideline they became aware of the problem of the linguistic fragmentation of Europe and the need to help to do something about that.

I was in fairly close contact at that time with DG XIII and when the question came up of what to do, and where to get a machine translation system for the European languages, not for Russian and English — they were not interested in Russian and English — we looked around, we looked at the research projects, and there were very few around in Europe, which might have served as the basis for a European development. After fairly painful discussions it was decided that none of them was mature for such a development, that the only candidate which could serve as a basis was a successor or in part parallel development to the Georgetown system which had started also in Georgetown but then took a different line, which was called SYSTRAN, and which had been sponsored by the foreign technology division of the United States Air Force ...

My main system of course was Russian to English, but they had made a number of prototypes or mock-up systems showing that the same approach could be used for other languages.

GK: *So, DGXIII got in SYSTRAN and that's fairly well- documented. You were involved in those discussions and evaluations?*

Perschke: I was involved in the discussions and evaluations and maybe it was my final advice which led to the decision to take SYSTRAN rather than the European systems.

GK: *Then we must have come to that meeting in February 1978, which made the crucial decisions.*

Perschke: We were almost there at the end of '75. I think technically it was a good decision and the same branch of this department looks after Eurotra, and at the other end of the corridor looks after SYSTRAN. SYSTRAN has been developed for quite a number

more language pairs, it has developed relatively big dictionaries and is being used by our translation services yet on an experimental or trial basis...

Politically this decision was not very popular in Europe. The Europeans had financed, not very much, but had financed some research in machine translation, particularly the Germans and the French. They realised the level of financing had not been enough to lead the research to a maturity which could then be used as the basis for product development. They didn't like the idea that European money should go to an American product but they also realised that if nothing happened the same thing would happen again and again. If it was not the Americans then it would be the Japanese, for the next generation. The decision to take SYSTRAN actually led, in the next two years or so, to, at first the idea, but then the decision, to start a European programme or project which should prepare for the more medium- or long-term future which would not depend on what would happen in Japan or the United States.

GK: *It's often written as if Eurotra came straight out of that decision but if you read up on the subject there was some collaboration between GETA and SUSY, there was a Leibnitz project? Did that have any influence?*

Perschke: It was never a real project. There was a a so-called Leibnitz group with GETA, SUSY at Saarbrücken, and Montreal, and a few other people. It was a very loose grouping which was discussing whether, or on what, or how one could cooperate, what one could do in the future and they all had been so long in isolation actually they were not prepared. The moving force in this Leibnitz group was Bernard Vauquois, whose dream I think was to convince the others to follow the GETA line, to contribute to the path of development of the GETA line and the others were not very keen on this. They wanted their independence. They discussed for a few years and one day they forgot to establish the date of the next meeting and everything fell apart. Nobody heard any more about the Leibnitz group.

GK: *The fact that you came up with a transfer system, to some extent like GETA and SUSY, was dictated more by the multilingual requirement than what had gone before that?*

Perschke: That is right. And in the first phase of the preparation of Eurotra, both GETA and SUSY participated very actively in the preparation. That however had to follow a line that neither SUSY nor GETA would become the nucleus of the new system. We didn't believe it was advanced enough and we had in addition the psychological and political constraints. If Eurotra were ever identified as the French or the German system then the others just wouldn't play the game. We wanted a European system. SUSY played the game; GETA a little less. For a long time they tried to push the design and preparation in that direction. It was that period influenced by the fact that the French Government was contemplating the launching of a national project in machine translation, to pick up the results of GETA and make an operational system. They considered Eurotra as competition.

GK: *Anyway you had a working group and then you had a Council confirmation on a three-phase programme, the first framework.*

Perschke: As far as the Community was concerned the only real competence in machine translation was in Ispra, not here in Luxembourg. Although I participated in the preparation in a sense for this meeting in '78, there were some discussions already taking place. I had participated in these discussions... The idea was that we should participate actively in the preparation and execution of this project but priorities, and the organisation of Ispra, had changed and we did not manage to convince the hierarchy of Ispra to commit themselves and participate in this programme preparation. So eventually I moved from Ispra to Luxembourg.

GK: *What year would that be?*

Perschke: This was mid '79. The preparation took place in parallel along two channels. Firstly, the technical preparations, to make up our idea what a new European system should look like, and for this reason it was called the Eurotra Coordination Group, at first with the participation of only a few countries. GETA was there for France, also SUSY, the University of Manchester, Essex had a project, a small university-type project, with real government funding, just the interest of a few researchers, and we used a research group in Switzerland, ISSCO, as the coordinators of this effort mainly ultimately because of considerations of neutrality...

GK: *So we have the Eurotra Coordinating Group and then at that time .....*

Perschke: We sort of made the blueprint of a possible project and system by mid '80. A very, very tentative blueprint already in '79 and by mid '80 we had a fairly precise idea. It was very very good.

We started thinking of proposing a research programme to the Community. It took a couple of years till June '80 to convince first the Director General and then the other Directors General, then the Commissioners, the decision-making body, that it was a good project and that a proposal should be made and such a proposal was so new and so unexpected that the Council, Parliament and so on, it took them two years from November '82 to come to a decision; and then all the problems like

the French problem and the German problem and that kind of thing all came up again and were all discussed and rediscussed, and in particular the perception of what such a programme or project should be was very different and it ranged from straightforward development of an MT — SYSTRAN, let's say — to fundamental research and we tried to do something with this.

I don't think it would have been reasonable to have taken the technology as it was and develop something which would have been somewhere between us and maybe a little beyond the system Ariane. Nor would it be reasonable to just continue almost free academic research under the guise of a machine translation programme.

We agreed, — it was actually a political decision — that Eurotra should not be a competitor to the potential generation of systems based on projects like GETA or SUSY, or the Canadian project TAUM, but it should prepare the ground for the generation after and to be more practical. The agreement was that the outcome was to be a prototype which would show the feasibility of a given approach to machine translation.

GK: *So you had a three phase programme, I believe.*

Perschke: In 1982, because of all these discussions and also because of the desire of the Council to have very close control over what happened the programme was sub-divided into three phases: a preparatory phase of two years which had a very small budget authorisation — two million ECU; a phase of basic and applied linguistic research into which a sort of very small scale prototype, a corpus-based prototype, was planned.

GK: *That's the 2000 word prototype?*

Perschke: Yes. And the third phase, which was of stabilisation and evaluation, expansion of vocabulary, which should move from a consultative approach to a more general approach, the vocabulary should be extended to approximately 20,000 words.

GK: *According to some articles, that 20,000 vocabulary includes 15,000 terms from telecommunications — that doesn't seem to be in any of the recent papers — is that idea dropped?*

Perschke: No, it’s not dropped. We are now approaching the end. I am not sure if we will get the 20,000 fully by the end of the project, but we will get reasonably close, and we actually chose telecommunications because of our Director General here was also very closely involved, ... and also because looking through possible sources of multilingual terminology, telecommunications would be good to choose because the EC's own terminology databank, Eurodicautom, had already, I think, about 12,000 terms in telecommunications in the nine languages. You could take over most of it but it had to be checked and one of the possible reasons for delay was that there was very little standardisation effort going on in telecommunication terminology ... Eurodicautom has added equivalents in other languages having taken a sample and asked people to have a look at as many extracts from the other languages. They advised us to review these equivalents or have them reviewed by experts thoroughly before the translations should be included in the dictionary.

GK: *And that has delayed matters a little?....*

Perschke: This created an additional complication because we have to look for experts, and negotiate with PTTs, who are not always very eager to divert their own experts, to what they think in general is something of a joke.

But 20,000 is an indication and it does not really make a great difference whether it is 20,000, 25,000 or 15,000 but it will be somewhere in the range of 15,000 to 20,000.

GK: *Will you be able to demonstrate the prototype with the expanded lexicon?*

Perschke: Yes.

GK: *When will this be?*

Perschke: Our estimation is at the end of the second half, from say September or so, of this year, maybe October. We have a last round of actual implementation which will finish at the end of August.

GK: *So you're more or less on schedule then?*

Perschke: More or less, yes.

GK: *It will surprise a lot of people who thought you were a long way behind schedule.*

Perschke: We couldn't achieve full synchronisation of all nine languages. You know another two languages have been added half-way, with the extension of the Community to Spain and Portugal — they started work only in 1987. In Greece we found a more or less green field situation, nothing like this had happened before...

GK: *It has been one of your problems, the scattered nature of the research. You mentioned it at Munich but it was inevitable for political reasons, I presume?*

Perschke: Not only political reasons, it depended on the nature of the project. It would always have been easier to set up a project in one place with well-defined hierarchical structures, command lines and what have you. But given the situation in Europe in computational linguistics and machine translation, it might have just meant a total brain-drain from all the countries to get to the critical mass in the centre. What, maybe, in the short-term, might have been easier perhaps would have been wrong for computational linguistics in Europe and for our political objectives and one of our objectives was to stimulate the development of expertise in this field, which was not there in a number of countries, and was also fairly thinly spread in the big countries. Another objective was to bring the Community to stimulate cooperation.

GK: *You have about 150 people involved, is that true? Some reports say 100, some say 150.*

Perschke: It is about 150 people, I would say.

GK: *And about 12 here, in Luxembourg?*

Perschke: We have right now a team of ten translators who are on loan from the translation services and six to seven people who are actually on the staff.

GK: *And in '87 you had this report that evaluated your progress so far; did the conclusions change your direction at all?*

Perschke: Somewhat, yes.

GK: *In what ways would you say?*

Perschke: I should give you first of all a bit of information about the background of the evaluation. In 1986 Spain and Portugal joined the project.

Because of the consequences of adding two more teams and two more languages we came out with the proposal for the inclusion of Spain and Portugal in the programme, and an adjustment of the timetables, and a mid-term revision of the programme.

Both the Council and the European Parliament agreed to this extension but they both asked for an in-depth evaluation of the programme before we would move to the third stage. This led to the famous Pannenborg report.

Well the conclusions were that both the pedagogic and political objectives had been maybe achieved better than one could have hoped at the beginning. It also said scientifically that it was very interesting and there had also been progress, but as far as actual implementation was concerned this was not terribly brilliant.

GK: *But I think the report was not as critical as many people expected at the time. Would that be fair to say?*

Perschke: I don't think the report was complacent but Eurotra is not as bad as some people go around saying... Eurotra has generated a lot of emotion. Many people react emotionally saying that it is brilliant or it is very bad depending on where they stand without actual in-depth knowledge of what the project is about and where it is.

GK: *If we could take up some of the criticisms. I think the Pannenborg report did seem to suggest that you concentrated too much on research and didn't keep the end idea of an operational system sufficiently in view.*

Perschke: They started criticising that the Council's decision of '82 did not specify what the end result ought to be; ...and you must not forget that the assessment panel was dominated by two fairly hard-nosed industrialists, Pannenborg and Danzin, a very well-known figure in French computing.

They emphasised the industrial aspect of machine translation, or of natural language processing in general, and as a matter of fact the interpretation which we get of the report, ...is that Eurotra has succeeded in carrying forward the research line. I say it is doubtful whether Eurotra was ever intended to go also into the industrial area.

Of course, again there is scope for interpretation: how, where the borderline between being competitive and research and product development is, but, the more industrial aspect of machine translation was never defined in the Council decision, it was not really aimed at by the mainstream Eurotra activities.

GK: *And in fact the report does seem to suggest that you should get industrialists involved in phase three or certainly immediately after, but __*

Perschke: OK, but they also recommend that we should not divert resources from research to try — which would have been in any case inadequate — to do something industrial instead. The research is needed. There is also one thing to say. Clearly, we cooperate in the mainstream research, mostly at the university centres, and the university centres have a specific attitude to, and an understanding of, what research is and they have an understanding of what development is, which in any other branch of technology would be seen yet as slightly short of applied research. It is too biased towards basic research. The other point is to say that by training they are not really equipped for the development work. They consider it boring, they have no experience, no real systems thinking, in engineering terms, to develop a system where machine translation systems would profit an engineering enterprise.

GK: *I think it was also suggested in the report though not so clearly that there was too much concentration on the linguistic side and not enough on the computing side. Did you perhaps recruit too many linguistics experts in the universities?*

Perschke: Maybe we recruited too many computer scientists from universities, but I believe we were not ready to go out to industry because we had to clear up our ideas. We put quite an effort into devising the software of the system, but the core concentration and the main effort went into defining foreign languages, formalisms, and the support, the underlying machine which would make the formalism computable.

But *because of* the university tradition in computer science itself and particularly in computational linguistics, as soon as the software prototype actually started running — very slowly and there were a lot of limitations — the perception of the computer scientists was that their task was finished, and the need to take such a software prototype which in principle works, and make software out of it, is not something you can do in the university.

So what Pannenborg and company saw was, I think, the second or third version of a formalism with a prototype implementation done to see whether things worked. It was absolutely inadequate to use this package as the software support for linguistic developments and experimentation, mainly because of limitation of speed and capacity. It would just break down if you overloaded it with texts or too-large grammars or too-large dictionaries — it just would not be able to hold them.

And this coincided with the period where the linguists actually started implementing. This was the second year of the second phase, we were aiming at the small-scale linguistic prototype and the work was actually pretty hampered by lack of software support. This is a problem which maybe had been underestimated at the beginning. ..

GK: *What will happen now? You will come up to the end of this year, demonstrate a prototype, what happens then? Are you negotiating a new framework programme or are you going to throw it out to industrial use?*

Perschke: We don't think that what we have is mature enough to be thrown out for industrial development nor is European industry mature enough to take up such a thing and invest in the infrastructure and they have been warned by somebody that they have to prepare for industrial involvement themselves, but don't expect miracles. Industry is not aware of the importance of languages, not aware of the economic potential in the long run and in any case the lead time before it gives profits is too long. So start preparing but don't expect miracles! They won't rush in and take over.

GK: *Will the prototype be publicly demonstrated?*

Perschke: Yes, we intend to make a public presentation, as we did at the MT summit last year. We shall make a presentation at COLING. We are preparing for this.

We are preparing now for the future in two stages; the short term and the medium term.

It so happens that Eurotra is now part of the second framework programme which is officially dated 1987-1991, but which has a fairly big overhang into 1992. We cannot commit money for a part of the money of the programme before 1992.

In this framework programme, ten million ECU are earmarked for the immediate follow-up of the Eurotra programme.

We have made a presentation, we presented the decision proposal to Council and Parliament.

We have started preparing for this transition, the first two years from 1991-92. Ten million is not a big sum to be divided between nine languages ...

We now believe that the general and theoretical framework of Eurotra is fairly well stabilised...

The internals of the dictionaries are defined by the most traditional machine, the environment, but the problem is the acquisition of dictionaries for machine translation and natural language processing; we would like to achieve a situation in which we can take existing dictionaries, SYSTRAN, for example, which has developed large dictionaries, or dictionaries developed for human consumption. And this last stage is to prepare for a project in the next phase of what we call re-usability of flexible resources and work towards standardisation of lexical or terminological data.

We are right now negotiating or selecting the potential contractors for these studies and negotiating the terms for going on the ground floor. We got quite a good reaction and I think we shall manage to get lots of industrial and scientific expertise together.

GK: *But you mention these are core aspects on which people will work. Will it all one day come together — and presumably we are talking about say 2000 — and Eurotra be the system which will solve the European multi-lingual problem, or is it something which has just been a stimulus and a catalyst? Will it be the Eurotra system?*

Perschke: I don't think the Commission can do all the job. Right now for a while, the stimulus for the Community programme is needed.

GK: *I gather from what you've been quoted as saying before that one of the disappointments of Eurotra was that a large element was entrusted to the Member States, but there was a certain impetus from the European Communities which was never matched except in one or two countries and this held you back in the early stages. Would that be true, people were slow to sign even the contracts?*

Perschke: Yes, it was true, yes. And every country has different reasons for being late, or slow.

GK: *So what we're saying is it's really only the European Commission which is aware of the scale and importance of the problems, not industrialists or individual countries.*

Perschke: The situation has very much improved in the individual countries. I can observe that one of the indicators is the interest both Council and the European Parliament and the press, in the attention they give to Eurotra.

To give an example, the Pannenborg report became available at the end of 1987. The second stage was officially finishing at mid '88 and due to a number of circumstances we are only able to make a proposal from the conditions that we had in the Spring of 1988. I think it was actually transmitted to Council on 1st June 1988. We had the decision on the transition on time, — it was great, maybe it can go in the record books.

We have discussed now the timetables for the approval of the transition programme and it should be a programme which starts in 1991; it should be approved formally by November. The proposal was presented in December, and we have created also conditions that the linguistic aspects are given a more prominent place in the new framework programme which hasn't been yet formally approved, but for which we are already preparing a programme. But I think I should pursue it in order: first the transition programme and then the new framework programme.

GK: *There's no part of the Eurotra programme which would bring in any discourse analysis or any hypertext; these are all things brought in since Eurotra defined the structure?*

Perschke: We tried to develop more and more semantic text to linguistic analysis for Eurotra. What we are considering of more importance or priority is to learn how to use non-linguistic knowledge in the linguistic environment.

GK: *So that's almost artificial intelligence?*

Perschke: There is a movement or a convergence with certain aspects of artificial intelligence. And this convergence is needed.

GK: *Because I thought the whole way Eurotra was structured with high level of formalism, and the analysis done very much at the source or target language, and not in the transfer, made it very difficult to bring in non-text information, or to access it. But are there going to be ways you could bring it in?*

Perschke: We have started already in the third phase in a limited manner

treating the terminology of the field of telecommunication. The terminology of the subject is actually the set of concepts which are relevant to the subject. And that's not only a set of concepts but it contains two types of pragmatic relations; one is of hierarchy between concepts and the other, things like instruments or tools and their function. .. So what we try, on the one hand to see, is basically the concept. So for the terminology we don't do any transfer, but we map the notations in each of the languages on the concept and then from the concept of the notation in the other language, so that this is already an approach to a potential dictionary.

But furthermore, we describe this set of pragmatic relations and link them to linguistic expressions, ... the station which transmits or the antenna, the size of the channel, the frequency, the power, the type of signal which is being transmitted and so on.

This is still experimental but this is a sort of connection to — call it a knowledge base.

GK: *Yes, because some of the criticisms made, of course, are that you created an obsolete system where there's no element for interactive, or there's no introduction of AI, but what you're saying is that all this is being looked into and brought in insofar as you can map it...*

Perschke: It's at a fairly tentative stage yet and I doubt we can give such a knowledge base within the formalism which was invented for linguistic data, linguistic knowledge, but I believe this is a very, very important aspect because you use only linguistic knowledge.

It never seems to be any normal context. You find out with any sentence, many more interpretations which one just excludes because it doesn't fit the expectations, and the over-generation, the resolution of ambiguities, is one of the crucial aspects in the nature of language processing, of machine translation too.