ACTIVITY REPORT 1991 – 1993 FOR EUROTRA-NL

Steven Krauwer

January 7, 1994

Contents

Introduction

This Activity Report of the NL Language Group covers the Eurotra Transition Phase 1991–92.

The report is structured as follows:

The first section contains some factual data concerning signature of contracts, financial provisions, and composition of the team.

The next section gives an overview of the Programme of Work for this period, and the research and implementation results obtained.

Given the scope of this report no detailed information on the content of the various research activities is provided, and the interested reader is referred to the various intermediate and final reports submitted to the Commission.

Section 3 is dedicated to other activities, and especially to those oriented towards the future.

Section 4 contains some concluding remarks.

1 Administrative Data

1.1 Contractual Arrangements

The contract for the transition programme was signed on 11 July 1991. The overall amount available for the execution of the work was 530 kecu, to be contributed by the CEC (318 kecu), and by NBBI, the Dutch funding agency (212 kecu). In addition an amount of 45 kecu was made available for grants. Work started on 1 January 1991, and ended 30 June 1993.

1.2 The Team

After the conclusion of the first Eurotra programme the team was reduced from ca 27 fte (35 individuals) to ca 17 fte (23 individuals). At that time members of the team were working on the following projects: Eurotra 91-92 (5 fte), Eurotra-LEXIC (part of the old Eurotra programme, 5 fte), Eurotra-GRAMMAR (a companion project funded by the Dutch government, 3 fte), ET-10 preparation (special subsidy from NBBI, 2 fte), and general supervision and support (2 fte).

After the first two quarters of 1991 the LEXIC and GRAMMAR projects came to an end (although some activities continued until the end of the year). This caused another reduction of the team to 14 people. In the meantime new projects have started (ET10, LRE) , and it looks as if the size of the team is stabilizing at the level of 12-14 people.

Below an overview of the composition of the team during 1991-92, the first two quarters of 1993, and a projection for the third quarter of 1993.

year:	1991				1992				1993		
quarter:	1	2	3	4	1	2	3	4	1	2	3
Bakker den	s	s	s								P
Bloksma	l	l	l	l	c	c	c	c	c		U
Bolhuis, van	l	l									P
Buenen				s	s	s	s	s	s	s	s
Dorrepaal	e	e	e	e	e	e	e	e	d	d	d
Eijk, van der	l	l	l	l	e	e					P
Florenza	e	e	e	e	e	e	e	e			U
Gaalen, van	e	e	e	e							P
Geilen	l	l									P
Groos	g	g									U
Heylen	e	e	e	e	c	c	c	c	c	c	c
Herklots	l	l									P
Hoekstra	e	e	e	e	e	e	e	e	n	n	n
Kamperman	g	g									A
Kester	g	g	x	x	x	x	x	x	x	x	x
Kraan, van der					e	e	e	e	r	r	r
Krauwer	s	s	s	s	s	s	s	s	s	s	s
Mineur								t	t	t	t
Ming	l	l									P
Munster, van	l	l									U
Pohlmann	e	e	e	e	e	e	e	e	e	e	U
Ruessink	e	e	e	e	e	e	e	e	r	r	r
Schenk					c	c	c	c	c	x	x
Steenbakkers											t
Tombe, des	s	s	s	s	s	s	s	s	s	s	s
Vercouteren	s	s									P
Weerden, van	s	s	s	s	s	s	s	s	s	s	s
Wouden, van der	e	e	e	e							A
(vacant)										b	b
(vacant)											v
(vacant)											v

abbreviations:

b: Robustness Project (Int. Coop.)
c: Collocations (ET10)
d: Discourse (LRE)
e: Eurotra 1991-92
g: Eurotra-GRAMMAR
l: Eurotra-LEXIC
n: NP-Interpretation (NBBI)
r: Reusable Grammar (LRE)
s: Supervision, management and support
t: Eurotra Grant
v: Text Retrieval (NBBI)
x: Local projects (not STT)
A: Academia
P: Private Sector
U: Unemployed or unknown

2 Goals and Achievements

The full NL/B programme of work for 1991-92 was presented in KOM 55222 (10-04-91), and modified in KOM 136118 (15-10-91) and KOM 33047 (10-12-92). Both the original programme and the modifications were approved by the LG.

Below we will go through the programme, and report on the achievements under the various headings.

2.1 Revision, Extension and Testing of Modules

A complete account of the implementation work done was given in Implementation Report Eurotra NL/B December 1991. This includes both the analysis extensions and revisions, and the analysis and synthesis revisions following from transfer extension and testing.

Implementation work done after the end of 1991 was limited to lexical extension (especially lexical harmonisation with transfer partners), debugging and maintenance.

2.1.1 Analysis

When looking at the programme of work one can say that the objectives have been met, although we would like to repeat a number of points mentioned in our Implementation Reports.

First of all a number of phenomena have only been implemented partially. The reasons for this can vary, e.g. in the case of impersonal passive where implementation will not be possible until more monolingual research work has been done on 'er', or in the case of coordination where neither the legislation nor linguistics in general offer any solution for the treatment of conjuncts with incompatible feature descriptions (which at least in Dutch are very common).

A second point is that it is very difficult to know whether a phenomenon is covered fully or only partially without having run the implementation through huge amounts of testing material. Not only new and unexpected variants of the phenomena may turn up, but in addition one may encounter cases of complex interaction between different phenomena, which may or may have not been taken into account by the legislation, and may or may not be implementable at all in our current framework.

As we all know time and space performance of the current ETS prototype do not allow for more than small scale superficial testing and we are well aware that successful completion of such testing should be interpreted as an indication that at least some known types of instances of the problem are dealt with properly, rather than a guarantee that the problem has been solved adequately in a general sense.

2.1.2 Transfer

The ES→NL transfer component is based on the deliveries made available to us by the source language group. It is complete in the sense that it covers the intersection of source

and target grammatical coverages. The quantitative coverage is ca 3600 entries, which we take to cover the source language transfer and demo text and the telecommunications corpus, although we have not actually checked this.

Because of practical problems we have not been able to perform thorough testing on the basis of the full dictionaries, but since the full dictionaries are well-formed, we trust that whenever the necessity arises arbitrary subsets can be extracted from the full dictionaries and loaded with the t-rules.

Text-to-text testing had to be abandoned, again for practical reasons. Most of the testing has been done with IS-objects provided by the source language groups.

2.1.3 Generation

A major difference between the final delivery and all previous ones is that we have abandoned the symmetry between analysis and generation.

We have done this for a number of reasons, but we would like to state here very explicitly that they are of a purely pragmatic nature. The desideratum to describe what is essentially the same body of linguistic knowledge only once remains, but given the stage of the project there are no longer compelling reasons to stick to this principle.

First of all the objective of this round of implementation work was not to extend our linguistic knowledge and expertise, but rather to provide those who do linguistic research with an appropriate test bed. Especially where we all know that new, and hopefully linguistically and computationally more adequate formalisms have been introduced (ALEP0, ALEP1), there is little point in making any more investments in types of linguistic elegance or sophistication that do not contribute to the performance of the components.

Secondly another objective of this last phase was to produce demo variants of the system, and again we felt that little would have been gained (if anything at all) by adding or even maintaining the sort linguistic sophistication embodied by symmetry of components.

Experience has shown that in a practical system analysis and generation require their own strategies in order to attain computational efficiency, which are not always compatible. In the past we have solved this problem (in cases where this was really necessary) with ad hoc fixes, but once it has been decided that a system is no more than a tool, there is little reason to maintain self-imposed constraints that have a negative effect on efficiency.

As a result the final grammatical delivery contained separate analysis and generation modules. Although they both derive from the same underlying symmetric modules, they have been refined and adjusted in different ways.

The separation of analysis and generation work has also led to a redistribution of monolingual tasks. Leuven was responsible for generation, and Utrecht was responsible for analysis and the monolingual dictionaries.

2.2 Research

2.3 Monolingual Research

Dutch Crossing Dependencies

The main result of the research on Dutch Crossing Dependencies is that the problem seems to be an artifact of traditional GB thinking, which has in certain respects influenced the Eurotra model. In modern grammar formalisms, such as HPSG, a number of solutions have been proposed. Since the ALEP formalisms to a large extent follow the ideas underlying HPSG thinking the DCD problem has been reduced to a selection problem between a number of alternatives. The actual choice cannot be made in isolation, and will have to depend on other linguistic properties of the grammar in which DCDs will have to be analyzed.

Dutch 'er'

The aim of this research was to suggest a possible implementation of the complete picture of 'er' plus related problems within the present E-framework. After the initial research phases it was decided to adopt HPSG as the linguistic basis for further research. The advantage of this reorientation was twofold. First of all this would lead to a treatment of 'er' that would not just be suitable for adoption by Eurotra, but rather for a much wider community. And secondly, since the ALEP formalism started to emerge it was felt that the HPSG orientation of this formalism would ensure relevance of the research results for future EC and other R&D programmes. Therefore the goal of the study was redefined as to work out an HPSG-mechanism that should account for the complex of facts involved in Dutch er-distribution.

The final report contains (i) a description of all aspects of HPSG that are relevant to the topic, (ii) an introduction of language particular constraints for Dutch HPSG, (iii) an extensive overview of all the relevant facts concerning the Dutch 'er' phenomenon, followed by (iv) an attempt to give an explanation of the observed facts.

ALEP experiments

During the period January – July 1993 small scale experimentation with ALEP0 was undertaken (after approval for this extension of the programme of work by the Liaison Group in December 1992). The experiments were carried out in close collaboration with researchers involved in ET10 Collocations, and LRE Reusable Grammar and Discourse. A report on these activities has been sent to the Commission.

2.4 Contrastive Research

The contrastive research on Scope, Determination, Negation, Quantification and Aktionsart has resulted in:

- A proposal for the representation of NPs, based on research on the NP specifier system. This proposal is an alternative for the current treatment of NPs in Eurotra, which is not satisfactory.

- A proposal for the representation of the scope relations between negation and (other) S-level modifiers, and of Neg-raising. This proposal is an extension of the current Eurotra legislation on negation.
- A proposal for the calculation of Aktionsart, which takes into account the interaction of Aktionsart with the verb on the one hand, and the different NP arguments on the other, and with negation. This proposal is an extension of the current Eurotra legislation on Aktionsart.

The proposals have been successfully implemented in Dutch, English, German and Spanish (analysis, transfer, and synthesis).

It is the intention to publish our the results of this research as a volume of the Studies in MT and NLP.

2.5 Transfer Research

Transfer research focussed on reversibility of transfer dictionaries, especially between Spanish and Dutch.

The final report was presented in the form of an article, and has been submitted to the 'International Journal of Lexicography'. It has been reviewed and is currently being revised.

The abstract is quoted below:

> The problem of organizing and classifying the variety of meanings that can be associated with a word in different contexts is a classical issue in lexicology and lexical semantics. A methodology is needed to determine whether two "candidate" senses should be grouped under a single heading, despite variation in meaning, or that two separate senses should be distinguished.
>
> Ambiguity tests provide a good starting point for such a classification, but must be applied with care. First, they cannot be applied to senses occurring in incompatible syntactic contexts. Second, the process of determining a set of discrete senses should not be confused with the polysemy/homonymy distinction. In this article we will outline and motivate a lexicographic methodology based on these assumptions. In addition, we will demonstrate in detail how the lexicographical description of a highly polysemous word, the Spanish verb *subir*, in the María Moliner dictionary would be reorganized by applying this methodology.

3 Other Activities

3.1 Training, Education

Two students have been given a Eurotra grant. The areas of research are 'HPSG and Categorial Grammar', and 'Evaluation of NLP Systems'.

The Utrecht Research Institute for Language and Speech (closely related to STT) has hosted EACL93, attracting more than 300 participants from 33 countries.

The policy to let researchers involved in externally funded projects (Eurotra, ET10, etc) teach specialized courses in the Computational Linguistics programme of the University has been maintained.

3.2 Future

During the execution of the 1991-92 programme the NL group has been actively (and fairly successfully) involved in various preparations for future actions.

ET10:

A number of proposals were prepared for ET10. The proposal 'Collocations and the Lexicalization of Semantic Operations' was accepted by the Commission, and work started on 1 January 1991. The actual work was concluded on 1 May 1993, but a small scale continuation (for revision and publication) is foreseen until the end of 1993. Participating Centres were STT (Utrecht, main contractor), University of Essex (Colchester), ISSCO (Geneva), ILC (Pisa), and Oxford University Press.

LRE–1:

Two projects submitted by STT were selected for LRE-1:

- 'Towards a Declarative Theory of Discourse', together with LTG (Edinburgh), University of Amsterdam, University of Clermont-Ferrand, and University of Essex.
- 'The Reusability of Grammatical Resources', together with LTG (Edinburgh), Universität des Saarlandes (Saarbrücken), and ITK (Tilburg).

Both projects have started 1 January 1993, and will have a duration of 2 years.

Furthermore STT participates in EAGLES.

LRE–2:

STT was partner in a number of project proposals. Two of the projects in which STT participates have reached the negotiation stage: Eurotext (corpora and corpus tools), end TEMAA (evaluation).

ESPRIT BRA:

STT participates in DYANA–2 ('Dynamic Interpretation of Natural Language'), together with ILLC (Amsterdam), CCS (Edinburgh), CIS (München), IMS (Stuttgart), ILF (Oslo) and SNS (Tübingen). Work started in October 1992, and will have a duration of 3 years.

DG XIII International Cooperation:

Together with CWARC (Montreal), and SITE (Paris), STT will work on the project 'Robustness: Combining Linguistic and Statistical Knowledge'. Work has started in May 1993, and will have a duration of 2 years.

NBBI:

NBBI, the funding agency for the NL Eurotra participation, has provided financial support for the following actions:

- Preparation of ET10 and LRE proposals (during 1991 and 1992).
- Linguistic research (project 'NP Interpretation', duration 1 January – 31 December 1993).
- Project 'Text Retrieval', in collaboration with Philips Research. Duration 2.5 years, start late 1993 or early 1994.

4 Concluding Remarks

A few points need to be mentioned here.

First of all it has to be said that in spite of all its shortcomings, the Eurotra programme has been a very stimulating enterprise, which has generated lots of beneficial effects.

When contrasting Eurotra with the follow-up programmes ET10 and LRE it is immediately clear that there are enormous differences.

Eurotra was oriented towards one specific goal, the design and production of a research prototype MT system. ET10 and LRE have no such specific goals. The clear advantage of the new situation is that it allows for exploratory activities, along different dimensions. The disadvantage is that there seems to be very little coherence within and across programmes.

Because of its long duration and the relative stability of the set of participants there seemed to be a reasonable balance between the resources needed to get involved and the research volume generated by the project (even if it has to be admitted that the communication overhead within the project was considerable – but paid for by the project).

ET10 and LRE have shown very clearly that the investment required to submit a proposal (several man months, with a 10% chance to succeed), and the resources needed to prepare the contracts (again a couple of man months) are enormous in comparison with the research volume generated (and paid for) by the project.