LS-GRAM
Final Review Report

Hans Ulrich Block,Lee Humphreys

1 Summary

1.1 The project

The project originally had the following objectives:

- creation of large-scale grammar resources for 9 languages
- creation of appropriate documentation (including a Rule Coding Manual)
- creation of a small MT demonstrator (2 language pairs)

These objectives were slightly revised during the mid-term review (see below).

The grammars were developed within the Alep framework, a grammar development environment designed to support unification grammars with typed features.

1.2 The review

The final review was based on demonstrations of analysis modules for 8 of the 9 languages - (Saarbruecken April 25-26 96). (Due to the absence of the French representative and some minor setup problems, it was not possible to demonstrate the French module.) Documentation (Grammar description and Test Results) was supplied for each analysis module.

1.3 Main findings

The English, German and Spanish grammars showed very substantial improvements in coverage since the mid-term review. The Extension grammars (Danish, Dutch, French, Greek, Italian, Portuguese) do not have the same level of coverage as the core grammars. Their development has benefitted from the Core language work.

Project documentation is of exemplary quality.

The project Coordination (Saarbruecken) appears to have been very effective and successful.

2 The project

According to the Technical Annexe of the project contract, the project had the following objectives

- creation of large-scale grammar resources for 9 languages
- creation of appropriate documentation (including a Rule Coding Manual)
- creation of a small MT demonstrator (2 language pairs)

The project plan envisaged the use of a corpus study in order to prioritize grammar development. Some emphasis was to be placed on treating of the complex phenomena of real texts (e.g. punctuation, complex coordination, long sentences), including so-called "messy details" (dates, proper names, numbers etc).

The tool used was ALEP - a grammar development environment designed to support typed feature structure grammars. The development of ALEP has been funded by the EC.

The project has two main components

- LS-GRAM Core (development of core grammars for German, Spanish, English)
- Extended LS-GRAM (extension to 6 further European languages)

At the time of the final review, Extended LS-GRAM was well under way. Most of these projects have about 12 pm left through to completion.

2.1 MidTerm Review

The mid-term review of the project was conducted by the present reviewers.

At the time, we found that, despite the best efforts of the partners, the three core grammars (English, Spanish, German) were rather less advanced than one might have hoped. We noted that this was largely attributable to the poor performance of Alep (clumsy interface, very poor parse speeds). However, the maintenance and support of the Alep platform had just been taken in charge by Cray Systems of Luxembourg and the partners were confident that substantial improvements would be forthcoming.

In view of the state of the grammars, we proposed a slight redefinition of project objectives. Instead of "creation of large-scale grammar resources" we suggested

> ... what ultimately will be delivered will be more like "extended demo" grammars i.e. much more extensive than toy grammars, but incapable of handling all the significant grammatical phenomena in the chosen text type. These grammars should be scientifically respectable i.e. avoid gross hacking just to handle one or other corpus sentence.

We also suggested that

- the MT demonstrator could be abandoned inorder to maximise resources for monolingual development.
- Any serious efforts on text handling should be abandoned (time-consuming and not now very relevant). Likewise for handling of numbers, dates and proper names.

3 The review

3.1 What we reviewed

The final review was based on demonstrations of analysis modules for 8 of the 9 languages - (Saarbruecken April 25-26 96). (Due to the absence of the French representative and some minor setup problems, it was not possible to demonstrate the French module.) Documentation (Grammar description and Test Results) was supplied for each analysis module.

The Test Material is based on economic texts (German, English, Spanish) - the corpus - together with test suites which systematically explore certain constructions.

In addition, we saw demonstrations of

- An HPSG analyser for Dutch based on Gregor Erbach's ProFit extension of Prolog. The lingware for this analyser was shown to be compilable into and hence runnable on Alep. (Utrecht)
- A Emacs-Lisp tool for extracting Alvey lexical data into the English Alep lexicon (Essex)
- A Perl tool for performing a similar type of extraction from the CELEX lexical database (Utrecht)
- Various text handling components partially or entirely developed in the context of the project
  - A Tcl/Tk-based interactive interface for grouping proper names (Essex)
  - An Awk-based text handler which allows (amongst other things) analysis of numbers and dates, with normalised forms being passed into the Alep analyser by means of ELDIF and specialised lifting rules (Copenhagen)
  - A Prolog-based tagger for German - MorPro. Experiments are underway to connect the output of this tagger to Alep, thus reducing the parse space.

3.2 Findings

The English, German and Spanish grammars showed very substantial improvements in coverage since the mid-term review. Sentences of realistic length (eg 20 words) and complexity can now be processed in a reasonable time.

Besides the use of a faster processor, the speed improvements are apparently due to

- continued efforts to design efficient grammars e.g. by the use of the Alep equivalent of rule packets, underspecification techniques etc. This effort has been led by the German group
- continued improvements in the performance of the Alep platform e.g. by more efficient handling of hypotheses in the parse. These improvements are due to work by Cray Systems.

The Extension grammars (Danish, Dutch, French, Greek, Italian, Portuguese) do not have the same level of coverage as the core grammars. In some cases their processing speed is lower. It is however obvious that the development of the Extension grammars has benefitted from the Core language work.

In general, the linguistic approaches taken in all the grammars are approximately similar, with approximately similar semantic representations.

In all cases the lexica used are rather small e.g. a few hundred words. The use of large lexica (e.g. ¿ 2000 words) was not explored in the project.

The number of parses for some grammars seemed lower than we might have expected given the problem of attachment ambiguities (PP, coordination, relative clause, etc). Perhaps some grammars contain some "built in" attachment heuristics.

A certain amount of effort has been invested in the treatment of so-called "messy details" i.e. text-handling issues. Whilst it is true that this is essential for a production NLP system, it is something of a distraction in the context of the present project.

As noted in the mid-term review, project documentation is of exemplary quality.

As also noted in the mid-term review, the project Coordination (Saarbruecken) appears to have been very effective and successful.

3.3 Conclusions

The reviewers were pleasantly surprised by the substantial progress which has been made since the mid-term review. It is clear that the objectives as redefined in the mid-term review have been more than adequately satisfied. The project participants have obviously worked very hard to overcome initial difficulties and should be congratulated on the resulte achieved.

Whilst none of the analysis modules were capable of analysing all the sentences within the chosen corpora, and further improvements are still needed in parse time, the capacities of the core modules suggest with reasonable further develpment the use HPSG grammars for high-quality analysis in certain non-time-critical production applications may be possible.

The Alep platform has greatly improved during the project period, both in terms of the user interface and parse performance. The recent introduction of the so-called 'record parser' has made very significant speed increases e.g. 4x. However, further speed increases are still needed.

Given that the development of each grammar has required about 2 person years, and assuming that the Extension grammars reach in due course the same performance and coverage levels as the core grammars, the project deliverable (9 linguistically comparable language grammars + documentation) represents good value for money. We note that the use of staggered development (Core + Extension) is probably a contributary factor to this cost-effectiveness.

4 Dissemination of Project Results

As noted in the mid-term review, we would like to see the project grammars + documentation + Alep bundled into downloadable files on an ftp site. There could be a subscription policy if restrictions are necessary for legal or economic reasons. A CDROM distribution may also be appropriate.

The preparation of such deliverables may require some modest portion of remaining project resources.

Care should be given to establishing appropriate Web links to the MLAP documentation set.

We note that the project has generated a number of conference papers.

5 Recommendations for the Future

A detailed technical comparison of the Dutch ProFIT-based analyser with Alep may help to identify some areas in which performance gains would be possible. (Both Alep and ProFIT compile typed feature structures into Prolog term representations.)

6 Recommendations

6.1 General

As ALEP is already freely available within EU we strongly recommend to also make the LSGRAM deliverables available on a server for remote access (i.e. ftp) and also on CD-ROM. The LSGRAM grammars should be included in the ALEP package.

It is our strong impression that free availability of ALEP and the LSGRAM results also outside of Europe would promote ALEPs attractivity.

As to our knowledge ALEP and the LSGRAM grammars is - in terms of delivered results - the largest project worldwide in the HPSG related research, such a free availability would strengthen Europe's reputation and influence in this area of research.

6.2 Strategic

In the mid term review we suggested that

> what ultimately will be delivered will be more like 'extended demo' grammars i.e. much more extensive than toy grammars, but incapable of handling all the significant grammatical phenomena in the chosen text type. These grammars should be scientifically respectable i.e. avoid gross hacking just to handle one or other corpus sentence.

Furthermore, we suggested that the MT-demonstrator should be dropped and instead as much time as possible should be forseen for analysis grammar development.

Compared to the mid term review we are now facing a completely new situation:

1. ALEP is now in much better shape than a year ago. It is fast enough for realistic grammar development and demonstrator applications.

2. Grammar development in LSGRAM has proceeded much faster than expected. The grammars of the core groups for German, English and Spanish are now in state that allows the use in a demonstrator application. The grammars of the extended groups for Portuguese, Greek, French, Dutch and Danish are of course less developed at the moment, but as these partners still have some months to go, we can expect that these grammars will also reach a state comparable to that of the Core groups.

Therefore, we would strongly recommend now that the originally planned MT-demonstrator should be built in a follow up project. As the demonstration gave us excellent information on the current status, we think that we can make the following very concrete recommendation for such a project.

6.2.1 Recommended follow up project

We propose to build a German to English or German to Spanish MT-demonstrator for a restricted domain s.a. PC-documentation, software documentation or the like completely on ALEP and the LSGRAM grammars. For such domain we expect that a lexicon of about 5000 basic terms and about 15000 domain specific terms would be needed. The texts should be real life texts provided by a third party.

In order to give the LSGRAM partners envolved a chance to prove that their MT-system does a better job than existing commercial systems we recommend putting strong emphasis on evaluation. Therefore we propose that another third party should be given the task to adapt the dictionary of a commercial MT system to the selected domain on the basics of the same text material as the LSGRAM approach. At the end of the project, both system will be evaluated on the basis of new unseen texts of the domain by professional translators according to rigorously defined evaluation criteria. In order to keep costs low we propose not to extend ALEP for robust analysis and disambiguation. Instead, evaluation criteria should take into account that the ALEP demonstrator will not be robust and might produce more than one translation in case of ambiguity.

Such a project could be done with a relatively small budget:

Work package	Person Months
Domain selection and data analysis	6
Analysis development	12
Analysis lexicon	12
Transfer lexicon	12
Tansfer rules	12
Generation lexicon	12
Improvement of Generation with ALEP	24
ALEP platform support/development	12
Domain lexicon for commercial system	12
Evaluation	3
Sum	117

In case that the ALEP approach wins the evaluation such a project could have a significant impact on the practical use of unification based systems. In case that the ALEP approach fails an investigation of the reasons would provide the commission with very well founded recommendations on where to go in MT in future framework programs.

7 Annexes

7.1 Profile and Background of the evaluators

7.1.1 Hans Ulrich Block

Hans Ulrich Block studied Romance and Germanic linguistics in Münster, Paris and Cologne. He holds a Master's degree in Germanic linguistics from the University Paris III (1977), a Master's degree in Romance linguistics from the University of Cologne (1978) and a Dr. degree from the University of Cologne (1984). All theses were on topics related to machine translation. He worked at the Institute of Technology in Aachen (1980-1982) on a project concerned with statistical text analysis, as an assistant professor in the German linguistics department of the University of Cologne (1982-1985). Since 1984 he is doing research on Natural Language Processing in the Corporate Research Department of Siemens AG in Munich. Since 1990 he is leading the NLP research group there. His main topics of interest are Parsing technology, Grammar, speech understanding and machine translation.

7.1.2 Lee Humphreys

Lee Humphreys is a Computational Linguist based at GSI-ERLI (France) with interests in Translation technology and Technical Documentation. He was previously involved in the design and development of translation tools at SITE-EUROLANG and in the EUROTRA project at the University of Essex.

7.2 List of relevant documents