

Learning Computational Grammars  
TMR Project Nr. ERBFMRXCT980237  
Final Report

John Nerbonne

December 19, 2002

## Summary

The Learning Computational Grammars (LCG) was a Training and Mobility of Researchers (TMR) project funded by the European Union's *Directorate Generale XII* which ran from April, 1998 through March, 2002, and involved seven sites — The Universities of Groningen, Tübingen, Antwerpen and Geneva, University College Dublin, SRI Cambridge and Xerox, Grenoble. The goals of the project were to investigate the application of techniques from machine learning to automatic language analysis, especially the recognition of simple phrases (“chunks”) in text. The project furthered the state of the art in this area substantially, accounting for ten of the world's best processing results in various subareas.

The project was organized in a SHARED TASK paradigm, in which groups focusing on various techniques all tackled the same problem. This allow groups to share data and supporting software, and thus work more efficiently. Furthermore, the shared tasks were consistently published openly, which stimulated participation beyond the group of LCG members. At the Lisbon, 2001 meeting, the majority of participants (seven of twelve) were non-LCGers!

Finally, LCG accounted for several years of training for young European scientists, and kept in close touch with industrial needs, seeking opportunities for rapidly involving technology. The project web site is still being maintained at <http://lcg-www.uia.ac.be/lcg/>, and the four annual reports as well as publications, software and data are available there.

# Contents

<b>1</b>	<b>Introduction</b>	<b>6</b>
<b>2</b>	<b>Scientific Results</b>	<b>7</b>
2.1	Stimulation of the Field . . . . .	7
<b>3</b>	<b>Training</b>	<b>8</b>
<b>4</b>	<b>Collaboration</b>	<b>8</b>
<b>5</b>	<b>Industrial Involvement</b>	<b>9</b>
<b>6</b>	<b>Site Reports</b>	<b>9</b>
6.1	Antwerp . . . . .	9
6.1.1	Postdoctoral Researcher: Erik Tjong Kim Sang . . . . .	9
6.1.2	Site Coordinator: Walter Daelemans . . . . .	10
6.1.3	Training Activities . . . . .	11
6.1.4	Industrial Involvements . . . . .	11
6.1.5	Collaborations . . . . .	11
6.2	SRI International, Cambridge, UK . . . . .	12
6.2.1	Introduction . . . . .	12
6.2.2	Postdoctoral Researcher: Anja Belz . . . . .	12
6.2.3	Predoctoral Researcher: Rob Koeling . . . . .	13
6.2.4	Site Coordinator: David Milward . . . . .	14
6.2.5	Other Researchers . . . . .	15
6.2.6	Project Related Activities at SRI . . . . .	16

6.2.7	Training Activities . . . . .	16
6.2.8	Industrial Involvement . . . . .	17
6.2.9	Collaboration . . . . .	17
6.2.10	Publications . . . . .	17
6.3	Dublin . . . . .	18
6.3.1	Postdoctoral Researcher James Hammerton . . . . .	18
6.3.2	Site Coordinator: Ronan Reilly . . . . .	20
6.3.3	Training . . . . .	21
6.3.4	Collaborations . . . . .	22
6.3.5	Other Activities at UCD . . . . .	22
6.4	Geneva . . . . .	22
6.4.1	Postdoctoral Researcher 1: Adelina Hild . . . . .	22
6.4.2	Predoctoral Researcher: Alexander Clark . . . . .	23
6.4.3	Postdoctoral Researcher 2: Franck Thollard . . . . .	25
6.4.4	Other LCG Researchers . . . . .	25
6.4.5	Site Coordinator: Susan Armstrong . . . . .	26
6.4.6	Other Researchers at ISSCO . . . . .	26
6.4.7	Training Activities . . . . .	26
6.4.8	Collaboration . . . . .	27
6.4.9	Industrial Involvement . . . . .	27
6.5	Xerox Research Centre Europe, Grenoble, France . . . . .	27
6.5.1	Postdoctoral Researcher: Nicola Cancedda . . . . .	27
6.5.2	Local Project Coordinators: Christer Samuelsson and Eric Gaussier . . . . .	28

6.5.3	Other Researchers . . . . .	28
6.5.4	Project Related Activities at XRCE . . . . .	28
6.5.5	Training Activities . . . . .	28
6.6	Groningen . . . . .	29
6.6.1	Postdoctoral Researcher: Miles Osborne . . . . .	29
6.6.2	Predocctoral Researcher 1: Stasinios Konstantopoulos . . . . .	30
6.6.3	Predocctoral Researcher 2: Susanne Schoof . . . . .	31
6.6.4	Network Coordinator: John Nerbonne . . . . .	31
6.6.5	Related Groningen researchers . . . . .	32
6.6.6	Training activities . . . . .	33
6.6.7	Industrial Involvement . . . . .	33
6.6.8	Collaboration . . . . .	33
6.7	Tübingen . . . . .	34
6.7.1	Postdoctoral Researcher 1: Hervé Déjean . . . . .	34
6.7.2	Postdoctoral Researcher 2: Franck Thollard . . . . .	35
6.7.3	Predocctoral Researcher 1: Alexander Clark . . . . .	36
6.7.4	Predocctoral Researcher 2: Yuval Krymolowski . . . . .	36
6.7.5	Predocctoral Researcher 3: Wouter Jansen from Groningen . . . . .	37
6.7.6	Site Coordinator: Dale Gerdemann . . . . .	37
6.7.7	Related Tübingen Researchers . . . . .	38
6.7.8	Training Activities . . . . .	39
6.7.9	Industrial Involvement . . . . .	39
6.7.10	Collaborations . . . . .	40



# 1 Introduction

The Learning Computational Grammars (LCG) was a Training and Mobility of Researchers (TMR) project funded by the European Union’s *Directorate Generale XII* which ran from April, 1998 through March, 2002, and involved seven sites — The Universities of Groningen, Tübingen, Antwerpen and Geneva, University College Dublin, SRI Cambridge and Xerox, Grenoble. The first five are universities and the last two are private companies.

The goals of the project were to investigate the application of techniques from machine learning to automatic language analysis, especially the recognition of simple phrases (“chunks”) in text:

[<sub>NP</sub> He ] [<sub>VP</sub> reckons ] [<sub>NP</sub> the current account deficit ] [<sub>VP</sub> will narrow ]  
[<sub>PP</sub> to ] [<sub>NP</sub> only £ 1.8 billion ] [<sub>PP</sub> in ] [<sub>NP</sub> September ] .

This example contains eight “chunks”, four NP chunks, two VP chunks and two PP chunks. In particular, the recognition of NP chunks is an essential step in the commercially interesting process of text classification for information retrieval, and in the automatic extraction of information give previously defined templates (“information extraction”).

During the course of the project, many techniques were applied, including Maximum Entropy, Instance-based (Memory-Based) Learning, Neural Networks, Explanation-Based Learning, Theory Refinement (Rule Induction), Inductive Logic Programming, Automata Induction, Genetic Algorithms, and Unsupervised Distributional Clustering. Finally, we were curious about the possibility of combining different techniques, including those from statistical and symbolic machine learning.

The project was responsible for several of the world’s best results on various publically defined learning tasks and was feature in an invited talk to the most important conference in this area, and also in a special issue in a leading journal.

To promote interaction, the project focused on a series of SHARED TASKS, which each group approached using different techniques. At various times in the project several of the world’s best processing results belonged to project members (see below). Due to the open structure of the project, it stimulated a great deal of work outside the group of funded partners (see § 2.1 below).

The project involved plenary meetings in Groningen (kick-off), Bergen, Cambridge, Dublin, Tübingen and Toulouse (final). A large number of people also attended the CoNLL meeting in Lisbon, 2000, where LCG sponsored a public comparison and competition of learning algorithms on a predefined task, as well. This functioned a bit as an informal meeting.

The project made use of nearly 220 postdoc months and 110 months of work by predoctoral researchers. Thirty-five senior researchers were not funded directly by the project, but attended meetings and interacted with the project researchers.

The current report summarizes the scientific results of the project, its contribution to training, the collaboration within the project, and the involvement of industry. We close with subreports from the seven sites involved in the postdoc network.

## 2 Scientific Results

A special issue of *Journal of Machine Learning Research* 2, March 2002, focused on applications of Machine Learning to the recognition and classification of basic grammatical structures in text (SHALLOW PARSING). Four of its seven papers arose from LCG:

**LCG General** “Introduction to Special Issue on Machine Learning Approaches to Shallow Parsing” James Hammerton (Dublin), Miles Osborne (Groningen), Susan Armstrong (Geneva), Walter Daelemans (Antwerp)

**Memory-Based Techniques** “Memory-Based Shallow Parsing” Erik F. Tjong Kim Sang (LCG Antwerp)

**Rule Abstraction** “Learning Rules and Their Exceptions” Herve Déjean (LCG Tübingen, then Xerox Grenoble)

**Maximum Entropy, Ensemble learning** “Shallow Parsing using Noisy and Non-Stationary Training Material” Miles Osborne (LCG Groningen)

Other scientific highlights included:

- Three of the world’s four best results in the on the recognition of simple noun phrases in text (Koeling, Tjong Kim Sang and Déjean, Bergen and Dublin, 1999);
- Four of the world’s seven best results in the so-called “text-chunking” task (Lisbon, 2000);
- Three of the world’s five best results in the clause recognition task (Toulouse, 2001);
- Eight LCG talks at the Conference on Natural Language Learning (CoNLL), held in conjunction with ACL 2001, where John Nerbonne, project coordinator, gave an invited address on the accomplishments of the LCG project. He held this talk with Erik Tjong Kim Sang and Hervé Déjean, using material from 10 junior researchers from LCG (Nerbonne, Belz, Cancedda, Déjean, Hammerton, Koeling, Konstantopoulos, Osborne, Thollard & Tjong Kim Sang 2001).

### 2.1 Stimulation of the Field

LCG organized its work around several SHARED TASKS involving the application of machine learning to natural language. The tasks all concerned the recognition of “shallow” syntactic

structure, e.g., “chunks” of various categories illustrated above. The tasks were publically defined, and essential material was made available — annotated data for training and software for the evaluation of results.

The shared-task organization meant that the network was focused not only by its scientific goal, the application and evaluation of machine-learning techniques as used to learn natural language syntax, and by the subarea of syntax chosen, NP syntax, but also by the use of shared training and test material, in this case material drawn from the Penn Treebank (above). The network scheduled three open workshops in which several external groups participated (Bergen 1999, Lisbon 2000, and Toulouse 2001), sharing data and test materials. In general about 50% of the groups participating in the task comparison were NOT LCG members. The opportunity to use data prepared by LCG, and the chance to compare results was so stimulating that these groups participated voluntarily.

### 3 Training

The LCG postdocs each worked under the supervision of a senior researcher, and we estimated that 10% of the postdoc’s time may be regarded as advanced training, providing 22.6 person-months of training. LCG predocs worked of necessity under much closer supervision, so that we estimate that 50% regarded as training, or 39.5 person-months. In general, researchers are at university sites provided an additional nine months each of training to graduate students and advanced undergraduates (see site reports) for an additional 45 person-months of training, and researchers at the industrial sites took or gave another one month of additional courses (2.5 person-months).

Special training was provided by Xerox, Grenoble, who conducted a course in the use of their finite-state tool kit for all of the LCG younger researchers. This accounted for roughly two person-months of training (including in the nine-person months which younger researchers document in the site reports).

We estimate the LCG contribution to training to be approximately the sum of these, i.e., 109.6 person-months. We have not included in this sum

### 4 Collaboration

There was extensive collaboration within the LCG project. Due to the shared-task paradigm, there was collaboration on task definition, data preparation and software for the evaluation for each of the three shared tasks. This material is publically available at the LCG web site, <http://lcg-www.uia.ac.be/lcg/>. In addition several LCG members made extensive visits to other laboratories in the network, including visits to Dublin and Tübingen for meetings, extensive visits to Antwerp from Dublin, Antwerp, and Tübingen; several visits to Groningen from Antwerp, Dublin, Tübingen, and Cambridge (one visit of five months); extensive travel between Tübingen and Geneva. These are documented in the annual reports

and in the site section below (§ 6).

## 5 Industrial Involvement

LCG included two industrial partners, Xerox, Grenoble, and SRI Cambridge, and it is fair to say that most of the time spent by researchers at these sites was devoted to research with an eye to practical application. In addition, Groningen explored an application of machine learning to language in cooperation with BSC, a local customer contact company; Geneva collaborated with Xerox and the World Intellectual Property Organization on terminology extraction.

## 6 Site Reports

This section contains the reports of the seven network sites, Antwerp, Cambridge, Dublin, Geneva, Grenoble, Groningen, and Tübingen, for the period April 1, 1998 – March 31, 2002.

### 6.1 Antwerp

#### 6.1.1 Postdoctoral Researcher: Erik Tjong Kim Sang

The research task of Erik Tjong Kim Sang in Antwerp was investigating the opportunities offered by memory-based learning for discovering syntactic analyses of natural language. Memory-based learning was already an established technique employed in Antwerp and its sister group ILK in Tilburg, The Netherlands, and some work on memory-based syntactic analysis was already in progress when Erik began as a postdoc (for example (Daelemans, Buchholz & Veenstra 1999)). He started by working on identifying basic nominal phrases, moved up to basic phrases in general and clauses, and ended with performing full parsing. He examined different methods for improving the performance of memory-based classifiers: cascading, system combination and feature selection. A unifying thread through Erik's time as a TMR postdoc was his involvement in the organization of the shared tasks of the yearly workshop on Computational Natural Language Learning (CoNLL). The shared tasks are still being used as benchmark tests for machine learning systems applied to natural language.

Erik's first work concerned using memory-based learning to identifying basic nominal phrases. He examined the influence of cascading and the representation of the output format of this task in a study together with Jorn Veenstra from the University of Tilburg in The Netherlands (Tjong Kim Sang & Veenstra 1999). Erik worked together with Miles Osborne from the University of Groningen in The Netherlands in organizing CoNLL-99 and its shared task which involved identifying embedded nominal phrases. The results of a study about

combining the results of systems using different output representations, both for basic and embedded nominal phrases, was published at NAACL-2000 (Tjong Kim Sang 2000a).

In 2000, Erik organized the CoNLL-2000 shared task, arbitrary phrase identification, together with Sabine Buchholz from the University of Tilburg in The Netherlands. He used his system combination method in this task and it finished third of eleven participating systems (Tjong Kim Sang & Buchholz 2000). He was the initiator of a cooperative study in which seven researchers from five different countries combined learning systems and together performed the best result for nominal phrase detection of that moment (Tjong Kim Sang, Daelemans, Déjean, Koeling, Krymolowski, Punyakanok & Roth 2000a). One year later, Erik applied memory-based learning to a more challenging task: identifying clauses. This was the shared task for CoNLL-2001, a task which he co-organized with Hervé Déjean from the University of Tübingen in Germany. Erik's system finished third of six participating system in this competition (Tjong Kim Sang & Déjean 2001).

At the end of 2001, Erik published the results of a study on using memory-based learners for building a full parser (Tjong Kim Sang 2001b). An overview of all his work as a TMR postdoc is presented in a paper in the Journal of Machine Learning Research (Tjong Kim Sang 2002). The main findings of this work are that memory-based learners are good in finding basic language structures, like text chunks, but that we have yet to find a configuration in which they are able to be competitive when dealing with hierarchical structures. In his three years as a TMR postdoc, Erik has published twelve conference papers, one journal paper and one book chapter.

### **6.1.2 Site Coordinator: Walter Daelemans**

During the period of the project, CNTS has strengthened its position as one of the main groups in Europe working on Machine Learning of Natural Language. Currently, 6 researchers (coordinator excluded) are working full-time in this area. Walter Daelemans attributes this in part to the contacts, interaction, and research activities within the LCG project.

Several additional projects in the general area of machine learning of language were acquired during this period. Apart from shallow parsing as an application area, memory-based learning algorithms were investigated also for grapheme-to-phoneme conversion, tagging, named-entity recognition, topic detection, summarization, prosody, and semantic disambiguation. Apart from memory-based learning, other machine learning methods were investigated as well, including rule induction and genetic algorithms (and more generally selectionist algorithms).

There has been a moderate interest from industry, culminating in two industry-funded projects, one funded by E-corporation, and one by a Lernout & Hauspie dependent research lab (CELE). Indirect industrial interest is evident from the fact that several companies have agreed to serve on advisory boards of some local research projects.

### 6.1.3 Training Activities

Erik's training activities in the four years of the project are summarized in the following table:

Time	Activity
0.45	Undergraduate teaching in year 1 (8 persons, 9 hours)
0.90	Undergraduate teaching in year 2 (8 persons, 18 hours)
0.75	Undergraduate teaching in year 3 (8 persons, 15 hours)
2.40	Perl course in year 2 (16 persons, 24 hours)
0.60	Master thesis supervision in year 2-3
0.02	Attending EACL'99 tutorial on Maximum Entropy
0.20	Attending LCG tutorial on FST in Grenoble
0.05	Attending ACL-2001 tutorial on Summarization
3.60	Training contribution of local senior researchers
8.97	

The network has generated a total of 8.97 person months of training activities over the three years that a postdoc was appointed at our site.

### 6.1.4 Industrial Involvements

Over the four project years, the industrial involvements of Erik consisted of teaching activities at our industrial partner S.A.I.L. This concerns about 40% of his undergraduate teaching activities as well as the Perl course he taught during the second project year. There have been research involvements of the local research group in the industry. These will be reported on in the section by the site administration.

### 6.1.5 Collaborations

Erik has also worked together with different people from the LCG network. He joined Miles Osborne (Groningen) in the organization of CoNLL-99. Rob Koeling (Cambridge) participated in a chunker combination experiment organized by Erik which resulted in a joint Coling-2000 paper. James Hammerton (Dublin) visited our local research group in June 2000 and Erik co-operated with him on his work on using memory-based techniques in connectionist systems (published in a CoNLL-2001 paper). Erik worked together with Adelina Ivanova (Geneva) on a linguistic analysis of the output of memory-based chunkers but this has not resulted in a publication. Hervé Déjean (Tübingen) participated both in Erik's combination experiment reported on at Coling-2000 and was the co-organizer of the CoNLL-2001 task together with Erik. Déjean's Tübingen colleague Sandra Kübler visited Antwerp in December 1999 for an introduction to the memory-based learning software TiMBL.

There is a close collaboration with the local research group in Antwerp and the ILK group of the University of Tilburg in The Netherlands. During his years in Antwerp, Erik has worked one day of the week in Tilburg where he benefited from being in the same environment as people that were working on similar research projects: Antal van den Bosch, Sabine Buchholz, Jorn Veenstra and Jakub Zavrel. His presence in Tilburg has led to two joint papers with people from that group.

## 6.2 SRI International, Cambridge, UK

### 6.2.1 Introduction

During the reporting period (April 1998 – March 2002) SRI employed a postdoc and a PhD student. This report presents a general overview of the network related activities at this site and specific reports for the postdoc, the PhD student, the local coordinator and others. An overview of the training activities concludes this section.

### 6.2.2 Postdoctoral Researcher: Anja Belz

Anja Belz did her doctoral research on the development of a formal method for phonotactic description and on practical learning techniques for the automatic construction of such descriptions, using a specially developed genetic algorithm for the automatic construction of finite-state automata that generalise over given phonological data samples. Other research interests include morphology, general automata theory, neural networks, comparison of natural language learning methodologies, and speech recognition.

As the first research subject on the LCG Project, Anja adapted a previously developed genetic algorithm for learning generalised finite-state grammars for NP-chunking. The overall result was that meaningful generalisation can be achieved, but that the cost of using a GA to construct and generalise grammars is too high for wide-coverage grammar development.

The term *treebank grammar* has come to mean PCFGs that are extracted directly from bracketed and annotated corpora. Anja investigated the performance of such treebank PCFGs on the task of parsing unseen test sentences, and the effect different methods of grammar reduction have on parsing performance.

As a follow-on project from the treebank grammar research, Anja investigated different sources of low-cost structural information directly derivable from treebanks, such as the identity of parent labels, grammatical role and the depth of embedding of rule applications, and how to incorporate such context into treebank PCFGs.

In a subproject, Anja adapted a structure-sensitive PCFG for the task of arbitrary chunking (as defined by Tjong Kim Sang for CoNLL-2000).

Anja's project on treebank grammars and probabilistic context-free grammars (PCFGs) with local structural context (LSC) has three main components: (i) methods for deriving treebank grammars directly from corpora; (ii) incorporation of different types of local structural context into grammars and testing the effect on parsing results; (iii) automatic optimisation of grammars for given parsing task in terms of grammar size and performance. During each of the three research stages corresponding to these components, grammars are tested on four syntactic parsing tasks: (i) full parsing, (ii) base NP chunking, (iii) text chunking, and (iv) NP recognition. Results for the first two stages, and preliminary results for the third were reported in a paper at *Corpus Linguistics* 2001 (Belz, 2001).

The main focus of Anja's final research project for TMR-LCG was the development of a new method for automatically constructing probabilistic grammars for a range of parsing tasks. The learning method, *Grammar Learning by Partition-Tree Search*, takes a base grammar and parsing task (in the form of a corpus of target parses), and constructs a probabilistic context-free grammar for the given parsing task.

Belz has applied the method to learning Local Structural Context Grammars for shallow and partial parsing tasks, including two of the shared TMR-LCG project tasks.

The aims of this research included (i) investigating the general usefulness of Local Structural Context for making parsing decisions, in particular the usefulness of a new type of LSC, the *Depth of Embedding of Phrases*, and (ii) looking at how well nonlexicalised systems can perform in comparison to lexicalised ones. With respect to these aims, results have shown that (i) selective use of LSC can drastically improve parsing performance on partial and complete parsing tasks, and that (ii) the non-lexicalised LSC grammars that were tested are not as good as the best lexicalised systems (although they come close on partial parsing tasks).

Belz plans to continue research in this area, and will next add a form of head-lexicalisation to LSC-PCFGs. This is to complete an ongoing investigation of the hypothesis that current best shallow parsing results can be improved if a selected amount of structural context is taken into account, i.e. if a limited amount of "non-shallow" analysis is carried out during parsing, in addition to lexicalisation.

### **6.2.3 Predoctoral Researcher: Rob Koeling**

Rob Koeling worked previously on the grammar of a natural language processing module of a spoken dialogue system. His PhD research (Groningen) used contextual (dialogue) knowledge to improve wordgraph parsing. He previously looked at knowledge based approaches, and investigated the use of statistical (maximum entropy) models to exploit information in system questions for parsing user utterances.

Rob's LCG research investigated the possibilities of applying the Maximum Entropy modelling techniques to the research tasks defined for the project.

As a first exercise, Rob performed experiments on base NP chunking using MaxEnt models.

Building on results from baseNP chunking experiments, Rob started experiments for shared task 1: Annotating sentences with parentheses marking NP boundaries. First results were given at the LCG Project Meeting (November 1999), where he presented “MaxEnt NP Chunking”. A project report describing method and results for both series of experiments appeared (Koeling 2000a).

Some research was done on feature selection and smoothing techniques for MaxEnt models. His publications include “Using Maximum Entropy Modelling for Contextual Interpretation of Answers”, TST Technical Report 99 (September 1999), Koeling, 2000a, and Tjong Kim Sang et al., 2000 (see below).

Rob gave a local presentation and introduced a paper at weekly LCG seminars. He also presented “A Maximum Entropy Model for adding context in a spoken dialogue system”, *Computational Linguistics in the Netherlands. CLIN X*, University of Utrecht (10 December 1999).

Rob visited LCG partners Groningen University and Tübingen University. He attended EACL’99 and CoNLL-99 in Bergen, as well as the LCG project meeting in Grenoble (May 2000) and the Gotalog conference in Gothenborg (May 2000). A talk was presented at the project meeting in Grenoble and a poster was presented at Gotalog. The poster presentation at Gotalog resulted in a published paper (Koeling, 2000a) and two more papers were published in conference proceedings (Tjong Kim Sang et al., 2000; Koeling, 2000b).

The former publication (Tjong Kim Sang et al. 2000) was a joint effort with several LCG project members (among others). Koeling (2000b) was his contribution to the shared task defined for the CoNLL 2001 workshop of finding arbitrary syntactic chunks in text.

Furthermore, he contributed to the shared task of arbitrary chunking as defined for the CoNLL-2000 workshop, and, as part of the spring and summer series of seminars at SRI, gave a Maximum Entropy tutorial for local LCG researchers and some interested members of the Cambridge NLP community.

Koeling’s involvement with LCG ended 1 August 2000. He continued at SRI on commercial contracts for one year, when he returned to Groningen to work on a commercial contract with LCG partner Groningen.

#### **6.2.4 Site Coordinator: David Milward**

Dr. David Milward, Coordinator David Milward was the local project coordinator. He has particular interests in applying the results of the project to improve the parsing components of text processing systems used at SRI. In the project he had responsibility for supervising both Rob Koeling’s and Anja Belz’s work for the LCG project. He attended the LCG kickoff meeting in Groningen, the Dublin LCG meeting, and local reading groups on machine learning. He hosted the Cambridge meeting, and participated in the local LCG paper reading and seminar series. He has been looking into the use of the maximum entropy approach to noun group chunking within the SRI Highlight Information Extraction engine.

### 6.2.5 Other Researchers

Dr Richard Sharman, Director (SRI), takes an active interest in Maximum Entropy techniques and was involved in overseeing Rob Koeling's work for the LCG project using ME techniques.

Dr Stephen Pulman, Principal Scientist (SRI) and Reader (University of Cambridge Computer Laboratory), played a central role in the LCG project during its preparatory stages, particularly in planning the Cambridge part of the project, and has attended meetings and related activities (e.g. those involving ILP). He remained actively involved, relating the work of local full-time LCG researchers to the work of researchers at the Computer Laboratory (especially David Abensour and Sylvia Knight) and to his own interests in ILP. SRI is a member of the EC ILP2 end-users club, and Stephen Pulman has been helping to keep that project and LCG in touch. He was took a particular interest in LCG activities concerning inductive logic programming. He invited Dr James Cussens to give a seminar on ILP at SRI.

Dr Briscoe, Lecturer (University of Cambridge Computer Laboratory), presented a talk at the Cambridge LCG project meeting, "Automatic acquisition of subcategorization classes from textual corpora". He attended LCG seminars, reading groups and invited talks, and improved the GR annotation of a test corpus (from the SUSANNE Corpus) to cover NP internal structure which was used for evaluation purposes by project members (<http://www.cogs.susx.ac.uk/lab/nlp/carroll/greval.html>).

Sylvia Knight, part-time researcher (SRI) and doctoral researcher and tutor (University of Cambridge Computer Laboratory), presented a talk at the Cambridge LCG project meeting ("Decision tree learning"), and was involved on a small scale through discussions and meetings, and introduced two papers to the LCG reading group.

Other members of SRI's NLP research group, including Dr Ted Briscoe, Dr Ian Lewin, Dr Richard Tucker, Sylvia Knight, Aline Villavicencio, Ben Waldron, and other doctoral researchers at the Computer Laboratory, have taken an active interest in the LCG project, participating in the local LCG seminars and reading group.

Invited speakers at SRI included James Cussens, "Introduction to Inductive Logic Programming" (April 2000), Dr Miles Osborne (former LCG postdoc in Groningen, now lecturer in Edinburgh) who presented two papers ("Estimation of Stochastic Attribute-Value Grammars using an Informative Sample", and "MDL-Based Learning"), Dr Mark Hepple (Sheffield) who gave a presentation on treebank grammar research in Sheffield, and Aline Villavicencio who presented her paper "The Acquisition of a Unification-Based Generalised Categorical Grammar."

### 6.2.6 Project Related Activities at SRI

LCG Project activities at SRI during the reporting period focused on research on shared project tasks and related grammar learning problems (for details see below). The two main project researchers published a total of five refereed papers in international conference proceedings. There were regular LCG project meetings, in the form of seminars, a reading group and invited talks. Project researchers also maintained extensive project-related web pages, and attended project-related conferences, workshops, seminars and meetings.

LCG project research by the full-time project researchers is reported in detail below. The LCG group at SRI organised a series of regular LCG project meetings the form of which varies between seminars presented by local researchers, reading groups and invited talks. During Spring 2000 a special seminar series on Statistical Methods for NLP and NLL took place. This consists of three parts, a reading group, tutorials on Maximum Entropy Modelling and a series of invited talks (<http://www.cam.sri.com/tmr/local-lcg-seminars/>).

During the reporting period, SRI LCG project researchers attended conferences and meetings, including EACL'99 in Bergen, the LCG project meetings in Bergen, Dublin, Grenoble, and Tübingen, Gotalog (May 2000), Coling-2000 and SIGPhon-2000 (July). They also presented talks locally and at project meetings on results achieved for LCG learning tasks (for details, see below).

### 6.2.7 Training Activities

Rob Koeling attended the tutorial Natural Language Learning with the Maximum Entropy Framework by Adwait Ratnaparkhi at EACL'99 and he gave a tutorial on Maximum Entropy modelling at fellow LCG partner Tübingen. He has visited LCG partner Groningen University and attended the XFST course at XRCE in Grenoble in May 2000. Rob Koeling also attended the LCG project meeting in Grenoble (May 2000), including the Finite-State Methods course. As part of the spring and summer series of seminars at SRI, Rob Koeling gave a Maximum Entropy tutorial for twelve local LCG researchers and some interested members of the Cambridge NLP community. He also gave this tutorial to six members of the Tübingen LCG group.

Apart from her research activities, Anja Belz designed and maintained local LCG web pages, attended EACL'99, CoNLL-99, and the Royal Society Meeting on Computers, Language and Speech, as well as administrating and organising the local LCG Seminar Series, which included reading group sessions, presentations by local researchers (including tutorials) and invited talks by researchers from other organisations. Anja presented the following talks: *Genetic Search Algorithms for NP Learning* (16 July 1999), TMR-LCG Project Presentation, SRI International, Cambridge, *Learning Finite-State Noun Phrase Grammars* (4 August 1999), Local LCG Seminars, SRI International, Cambridge, and *NP Learning with Treebank Grammars and Genetic Algorithms* (16 November 1999), TMR-LCG Project Meeting, University College Dublin. She also attended the LCG project meetings in Grenoble (May 2000) and Tübingen (February 2000), as well as COLING 2000 and SIGPhon 2000

in Luxembourg (July 2000). She presented talks at the two LCG project meetings, the local LCG seminar series, SIGPhon-2000, as well as giving two invited lectures at UCD, Dublin (May 12) and ITRI, Brighton (June 8). Two publications resulted from her research during the reporting period: Belz (2000) and Belz (2001).

The SRI group generated four person-months of additional training.

### **6.2.8 Industrial Involvement**

SRI Cambridge is a commercial partner whose financing comes exclusively from research and development contracts, and primarily from contracts with private industry. The LCG researchers were frequently in contact with industrial concerns in Cambridge.

### **6.2.9 Collaboration**

Rob Koeling combined inter-LCG collaboration with industrial involvement when he visited Groningen between Sept. and Dec. 2001, during which time he worked together with John Nerbonne, Gosse Bouma and Tanja Gaustad (Alfa-Informatica) on an automatic email classification application (see § 6.6.7 on Groningen's "Industrial Involvement"). He also visited Tübingen to give a mini-course on maximum-entropy modeling, and Antwerp in order to work on ensemble learning with Erik Tjong Kim Sang.

Collaboration between the local full-time LCG researchers and associated project members, in particular at the University of Cambridge Computer Laboratory, formed an important part of the Cambridge LCG activities. Members of the Computer Laboratory have been attending Local LCG Seminars and presenting talks, and a special interest group in Inductive Logic Programming (Stephen Muggleton, Stephen Pulman et al.), has been very active organising projects, seminars and meetings. There was further collaboration between ILP projects and LCG in the form of regular meetings and discussion groups.

Anja Belz collaborated on a small project with Prof Dr Gerald Gazdar (Sussex University) looking at the correlation between the frequency of occurrence of different phrases and their depth of embedding in parse trees.

### **6.2.10 Publications**

(Belz 2001*b*, Belz 2000, Koeling 2000*b*, Tjong Kim Sang, Daelemans, Déjean, Koeling, Krymolowski, Punyakanok & Roth 2000*b*, Koeling 2000*a*, Belz 2001*a*, Belz 2002)

## 6.3 Dublin

### 6.3.1 Postdoctoral Researcher James Hammerton

James worked on the project from January 1999 to December 2001 then left to take up a post at the University of Groningen. He applied neural networks (NNs) to noun-phrase (NP) bracketing and clause identification on the Wall Street Journal (WSJ) corpus. Most NN research in natural language processing (NLP) involved small and/or artificial data sets since NNs often require intensive training. Thus applying NNs to real-world data sets required pushing NN techniques to use larger amounts of more complex data than had normally been used. James investigated several NN technologies for this work as follows:

**Scaling up NN Representations for Structured Data** The (Simplified) Recursive Auto-Associative Memory — (S)RAAM (Callan & Palmer-Brown 1997) — is a NN representation for structured data and offered quick training. However the (S)RAAM scaled poorly, requiring vectors of over 500 elements to represent trees derived from only 50 sentences from the WSJ corpus. The Sequential Activation Retention and Decay NETwork — SARDNET (James & Miikkulainen 1995), a self-organising map (SOM) for sequences, was investigated to see if it would scale better than the (S)RAAM. It forms dense representations of sequences, e.g. a 100-unit SARDNET was able to represent 3115 sequences of part of speech (POS) tags of length  $< 20$  tags from the WSJ corpus. But when used as outputs for NNs, SARDNETs experienced noise disrupting the representations making it difficult to decode them. A method of making the SARDNET representations more robust is needed to solve this.

**Applying SARDSRN for NP Bracketing** The SARDSRN (Mayberry & Miikkulainen 1998) augments the simple recurrent network (SRN) with a SARDNET. Starting with 10 to 50 sentences of length  $< 10$ , experimentation eventually enabled the SARDSRN to be applied to all 752 of the sentences of length  $< 10$  words in the training set (sections 15 to 18 of the WSJ corpus). The network was tested on all 184 sentences of length  $< 10$  wd. in the testing corpus (section 20 of the WSJ). It processed 550 of the training sentences perfectly, yielding a training set fscore of 96.2 and a testing set fscore of 49.77, processing 29 of the test sentences perfectly. Training took 26 hours. Because the training was too intensive, an alternative approach was needed.

**LSTM for NP Bracketing and Clause Identification** A new NN architecture designed to retain information for arbitrary time-periods, called Long Short-Term Memory (LSTM) (Hochreiter & Schmidhuber 1997), was investigated. Trained similarly to the SARDSRN, a LSTM network learned to process 736 of the training sentences of length  $< 10$  words perfectly, achieving an fscore of 100 on the training set. On sentences of length  $< 10$  words in the testing set it processed 46 of them perfectly achieving an fscore of 62.79. The training took 19 hours. Whilst this was an improvement, the training was still too intensive.

	Train correct, fscore	Test correct, fscore
MBL 0-0	76 , 16.62	7 , 9.16
MBL 1-0	181 , 59.23	3 , 11.19
MBL 2-0	96 , 34.36	2 , 10.10
LSTM	736 , 98.88	46 , 51.56
SARDSRN	551 , 92.04	29 , 32.10
ErikMBL 1-0	n/a, 97.89	n/a, 65.92

Table 1: Comparison with MBL. Given here are the number of sentences correct and the fscore. These differ from those in earlier reports due to unbalanced brackets being taken into account.

Hidden Layer	Sentences used	Train fscore	Test fscore
$12 \times 4$	Length < 10 wd.	98.19	70.14
$6 \times 4$	Length < 20 wd.	75.44	72.01
$8 \times 4$	Length < 20 wd.	77.22	71.25

Table 2: Selection of results for NP bracketing on sentences from sections 15 to 18 (train) and section 20(test) of the WSJ corpus.

To test whether the poor generalisation was due to the small data set, a MBL system was implemented using the same inputs and outputs as the NNs and trained with increasing left context. See Table 1 for the results. The MBL system performed poorly, but this is probably due to the way the task was presented. “ErikMBL” in Table 1 is more representative of what MBL is capable of. This used LCG Antwerp partner Erik Tjong Kim Sang’s MBL system for NP bracketing (which processes sentences differently) which performed best, indicating room for improvement with the NNs.

A new way of presenting the task to the NNs was devised which resulted in faster training, and LSTM was applied to bracketing the NPs for all sentences of length < 10 wd., yielding a best NP fscore of 70.14 on the testing set. This approach was also applied to all sentences of length < 20 wd., yielding a best fscore of 72.04 in testing. See Table 2 for details. Using a similar approach, James participated in the CoNLL 2001 shared task (Hammerton 2001). He tackled Part 3 of the task involving finding all the clauses in a sentence. LSTM networks were trained on the first 1000 and 2000 of the training sentences, yielding clause fscores of 45.61 and 50.42 respectively on the testing set. The performance was worse than with other entries (range 62.77 to 78.6). Since the CoNLL workshop, LSTM was trained on the first 3000 sentences yielding an fscore of 52.14. See Table 3 for details. The conclusion is that whilst progress was made in scaling NNs to use large real-world data sets, there is still some way to go before they can be applied reliably to such data.

**JMLR paper** James produced a paper on his SARDSRN/LSTM work which he submitted to the JMLR Special Issue on Machine Learning Approaches to Shallow Parsing (see below). It was recommended that the paper be resubmitted as a regular JMLR paper.

Hidden Layer	Training data used	Train fscore	Development fscore	Test fscore
12 × 4	First 1000	64.78	49.62	45.61
12 × 4	First 2000	65.07	57.62	50.42
12 × 4	First 3000	64.20	58.33	52.14

Table 3: Selection of results for clause identification on WSJ corpus sections 15 to 18 (training), section 20(development) and section 21(test). The networks were trained on the first 1000, 2000 and 3000 sentences from the training set respectively.

**Other Talks and Conferences** Additionally, James gave several talks during the project. Three talks on his PhD work were given, one during the department’s AI seminar series, one at the “Cognitive Science for the New Millennium” conference on the 16th & 17th May 1999 and one was given to present a paper entitled “Holistic Symbol Processing”, (Hammerton 1999), to the AICS’99 conference, University College Cork, Cork, Ireland, September 1999.

He also gave two talks on using SARDNET as a representation at the “EmerNet: International Workshop on Emergent Neural Computational Architectures Based on Neuroscience”, September 1999, University of Edinburgh, UK and at the “Emergent Computing: Self Organising Systems – Future Prospects for Computing” workshop, October 1999, the Manchester Conference Centre, UMIST, Manchester, UK.

James also presented, as co-author, a paper (Heffernan & Hammerton 12–15 December, 2000) on Paul Heffernan’s final year project to the Intelligent Systems and Architectures (ISA 2000) conference, University of Wollongong, December 2000. Paul was unable to present this due to work commitments. Finally, James attended the CoNLL workshops from 1999 to 2001, giving a short talk on his LCG work at the LCG session of the CoNLL-99 workshop, and presenting two papers to the 2001 workshop, one on using LSTM for clause identification and one co-authored with LCG partner Erik Tjong Kim Sang on a hybrid SOM/MBL system (see below).

### 6.3.2 Site Coordinator: Ronan Reilly

Ronan organised the Cognitive Science for the New Millennium (16-17 May, 1999) conference which marked the launch of UCD’s MA/MSc in Cognitive Science and included talks by Professor James McClelland and Professor Paul Smolensky, acknowledged pioneers in cognitive science. He also directed the MA/MSc in Cognitive Science which involved contributions from various affiliates of the LCG project at UCD. In September, 1999, Ronan taught a course on connectionist approaches to grammar learning at the Tuebingen-Sofia Graduate Programme in Computational Linguistics and Represented Knowledge (CLaRK). He presented a paper on systematicity in grammar learning at the Pacific Rim NLP conference in Beijing, October 1999 and gave an invited talk in December 1999 at the Cognitive Science Colloquium, SUNY Buffalo, on the topic of language evolution and development. Within the Cognitive and Computational Neuroscience Centre (CCNC), Ronan was involved with research concentrating on the neural basis for language learning and processing. Much of this

work focussed on a theory of cortical development proposed by Reilly (2001) called Cortical Software Re-Use (CSRU). Ronan was also a PhD examiner for the University of Groningen and had several publications, (Reilly 1999, Reilly (in press), Akhtar & Reilly 2001, Kechadi & R.G. 2000, Nenova & Reilly 2001, Reilly 2000, Reilly & Mackey 2001). Ronan also gave several talks in this period:

- Cognitive Science: Models, problems, and prospects. Invited talk, Department of Linguistics, Chulalongkorn University, Bangkok, March 2001.
- Object assembly and language development, Workshop on the relationship between cognitive and language development, University of Chulalongkorn, Bangkok, Thailand, March 2001.
- Computational modelling of eye movements in reading, First Workshop on Eye-Tracking, Trinity College Dublin. Contributed talk. (May 2000).
- Cortical Software Re-Use: A framework for understanding mirror neurons, Mirror Neurons and the Evolution of Brain and Language, Hanse Wissenschaftskolleg, Delmenhorst, Contributed talk. (July 5-8 2000).
- Cortical computational building blocks. EmerNet: 3<sup>rd</sup>, International Workshop on Current Computational Architectures Integrating Neural Networks and Neuroscience, Durham, England. Contributed talk. (8-9 August, 2000).

At the end of 2001, Ronan left UCD to take up a new post as Head of the Computer Science Department at the National University of Ireland, Maynooth.

### 6.3.3 Training

James' training related activities included the following (estimates of the person months used are included in brackets):

- Proposing and co-supervising (with Ronan Reilly and Nick Kushmerick) two summer projects (4 person months) employing a student called Paul Heffernan and proposing and co-supervising a final year project again employing Paul Heffernan (4 person months).
- James attended the course Xerox ran on their Finite State Toolkit for NLP in Grenoble in June 2000. (0.25 person months).
- He attended some of the psychologically oriented courses of UCD's Cognitive Science MSc, during the 1999–2000 academic session. (1 person month).

James was thus involved in a total of 9.25 person months of training. Ronan's activities included roughly 4 person months training related to NLP/machine learning involving regular

meetings with James, supervision of several research students involved with NLP and/or machine learning projects and supervision of another postdoc, Dana Mackey, on a project on the neural basis of language learning.

### **6.3.4 Collaborations**

James visited the University of Antwerp, from the 28th May to the 9th July 2000, working with Erik Tjong Kim Sang on producing a hybrid NN/MBL system, employing a SOM to perform memory editing. Performance was close to that of MBL whilst using a smaller number of comparisons. A paper, co-authored with Erik, was presented at the CoNLL 2001 workshop held in Toulouse (Hammerton & Tjong Kim Sang 2001). He also co-edited, with Susan Armstrong, Walter Daelemans and Miles Osborne, the *Journal of Machine Learning Research*, 'Special Issue on Machine Learning Approaches to Shallow Parsing'. This is now available online via <http://www.jmlr.org/>. This has given James valuable experience of the process of editing a journal.

### **6.3.5 Other Activities at UCD**

Members of staff involved in the Cognitive Science MSc/MA included Fred Cummins, Julie Berndsen, Arthur Cater, Gregory O'Hare and Mark Keane. Dr Arthur Cater was involved in organising a series of computational linguistics seminars at UCD. Speakers were invited from a wide range of research groups, both in Europe and the US. Among the invited speakers was Anja Belz, the LCG post-doc based at SRI Cambridge. Dr Julie Berndsen gave an invited talk to the Royal Society in London on her work on finite state models of phonology. Dr Berndsen was also in contact with Anja Belz with a view to possible future collaboration. Fred Cummins provided James Hammerton with code for LSTM and had discussions with James about the use of LSTM and about the burdens placed on the network by different output representations. As of 2001, the main computational linguistic related activities at the Dept. of Computer Science in UCD are now conducted under the auspices of the recently constituted computational linguistics and speech technology (CLISTE) research group.

## **6.4 Geneva**

### **6.4.1 Postdoctoral Researcher 1: Adelina Hild**

Adelina joined LCG Geneva and began her work on the project in June, 1999. Her research focused on the acquisition and processing of NPs by humans and related investigation of recursive NPs with emphasis on nominal phrases headed by deverbal nominals.

This included

- Review of available experimental research on the acquisition and processing of phrase structures by humans, with specific emphasis on NP acquisition;
- Error analysis of the results from the NP chunking experiment carried out by Erik Tjong Kim Sang at the Antwerp site (Tjong Kim Sang and Veenstra, 1999). The aim of the analysis is to develop an annotation standard for automatic parsing for the grammatical structures that pose processing problems.
- Tagging the multilingual 'EU Parliamentary Debates' corpus and developing the German and English lexicons and morphology. This task was carried out in conjunction with other ISSCO members – Sabine Lehmann and Pierrette Bouillon – whose role has been instrumental in developing the linguistic modules for the three languages chosen for the corpus. To date, the English and German morphology have been updated and refined (the French morphology was in a ready-to-use state), which allowed to undertake the tagging of the respective parts of the corpus. During a visit to the LCG partners at University of Tübingen, Adelina met with the team working on the VERBMOBIL project to assess the possibility of syntactically annotating the corpus.
- She attended the workshop on Computational Natural Language Learning in the course of the EACL'99 conference, Bergen, Norway.
- Adelina gave a talk on error-analysis results from the above referenced machine-learning experiment at the Dublin meeting of the project participants, November, 1999.

Unfortunately, for reasons of illness she was unable to work for a period of 4 months. She left the project at the end of October 2000 due to personal reasons.

#### **6.4.2 Predoctoral Researcher: Alexander Clark**

Alexander Clark started work on the project on October 1st, 2000, and worked until the end of the project on March 31st 2002. During some of this period he was also employed at the Tübingen site — see that section for details.

His undergraduate degree was in Mathematics from Trinity College, Cambridge, and he also has an M. Sc. in Knowledge-Based Systems from the University of Sussex. During the period of the project he completed his D. Phil. at the University of Sussex under the direction of Dr. Bill Keller. The title of his thesis was “Unsupervised Language Acquisition: Theory and Practice”. This was submitted in September 2001, and he successfully defended it in December 2001. The external examiner was Walter Daelemans, the LCG site coordinator at Antwerp.

Alexander Clark's LCG research has focussed on unsupervised language acquisition. There are two main motivations for this: first, this may cast light on the way in which infant children learn their first language. In particular, it may be able to clarify the extent to which information present in the primary linguistic data available to the infant child is sufficient to allow the child to learn his or her native language. Secondly, given the easy availability

of enormous quantities of unannotated linguistic data, there are sound engineering reasons to explore the use of unsupervised or partially supervised techniques.

This research has covered a number of different levels of language including morphology and syntax.

In addition he has currently been working on using what is called the Fisher Kernel method, a new Machine Learning technique that appears to be well suited to a variety of different NLP tasks.

This research has resulted in a number of publications:

- Clark, Alexander (2001) Learning Morphology with Pair Hidden Markov Models *Proceedings of the Student Workshop at ACL 2001*, Toulouse (55–60).
- Clark, Alexander (2001) Inducing Stochastic Context-Free Grammars with Distributional Clustering *Proceedings of CoNLL 2001*, Toulouse (105–112).
- Clark, Alexander (2001) Partially Supervised Learning of Morphology with Stochastic Transducers *Proceedings of NLPRS 2001*, Tokyo (341–348).
- Clark, Alexander (2001) Unsupervised Language Acquisition : Theory and Practice, D. Phil. dissertation, University of Sussex
- Thollard, Franck and Clark, Alexander (2002) Apprentissage d'Automates Probabilistes Déterministes, *Proceedings of CAp2002*, (to appear)
- Clark, Alexander (2002) Memory-Based Learning of Morphology with Stochastic Transducers *Proceedings of ACL 2002*, Philadelphia (to appear)

In addition he gave various invited talks:

- Pair Hidden Markov Models and Morphology Acquisition, COGS NLCL Seminar Series, Sussex University, December 1st 2000.
- Unsupervised Induction of Probabilistic Context Free Grammars with Distributional Clustering, TMR-LCG Project Meeting, Tübingen University, February 21st 2001.
- Learning Finite State Transducers with Pair Hidden Markov Models, Xerox Research Center Europe, Grenoble, May 16th 2001.
- Learning Finite State Transducers with Pair Hidden Markov Models, IDIAP, Martigny, June 12th 2001.
- Formal methods in Natural Language Processing, EURISE, University of St. Etienne, France, January 24th 2002.
- Information Geometry and Memory-Based Learning: Morphology, CNTS, University of Antwerp, May 17th 2002.

### 6.4.3 Postdoctoral Researcher 2: Franck Thollard

Franck Thollard was employed by the Geneva site from September 2001 until February 2002. He first joined the Tübingen site and then came to Geneva in order to work with Alexander Clark. A fuller description of his research can be found in the Tübingen section of the report. Franck Thollard worked in two directions : on the one hand he adapted his grammatical inference technique to a shared task on Noun Phrase Chunking. This task was provided by some other LCG researchers (mainly Antwerp's Erik Tjong Kim Sang). On the other hand, he worked on improving grammatical inference algorithms and on learnability of probabilistic models.

Once in Geneva, Franck Thollard and Alexander Clark worked together on improving the NP-Chunker. This work improve the chunker up to 90 % of good answers. This work is under submission at the International Conference on Grammatical Inference (ICGI 2002). On the other hand, Franck Thollard and Alexander Clark worked on a more theoretical paper. The two works (Np-Chunker and theoretical one) are published at the French conference on Machine Learning (2002). The international version of the theoretical work is under submission at the ALT<sup>1</sup> 2002 conference.

Moreover, Franck Thollard started a collaboration with the EURISE Team (Saint Etienne, France). This lead to a paper at the European Conference on Machine Learning (2002). Franck Thollard also collaborated with the Master Thesis of the university of Saint Etienne and supervised a student (Toufik Boudellal) who collaborated with Sonia Halimi (University of Geneva) and Alexander Clark.

### 6.4.4 Other LCG Researchers

Toufik Boudellal joined the LCG network in November 2001. He aims at working on language modeling by way of probabilistic formal grammars. He is currently working on smoothing such models. Probabilistic models are useful in language modeling for speech recognition, spelling correction, information retrieval, and many other application areas. After a study of the literature, Toufik Boudellal formalised a new approach to smoothing probabilistic automata. He was supervised by Franck Thollard, another LCG member. Toufik Boudellal worked in collaboration with the University of Saint Etienne (France). The work done will be continued and will end up with Toufik Boudellal's Master Thesis.

Toufik Boudellal also worked in collaboration with Sonia Halimi (from ISSCO, Geneva). They worked together on automatic classification of arabic documents. They used word collocations. Toufik Boudellal also collaborated with Alexander Clark on Arabic morphology.

Celine Reynal worked on the ISSCO tagging tools applied to the *Le Monde* corpus. This entailed three different tasks: definition of segmentation rules, development of lexical re-

---

<sup>1</sup>ALT stands for Algorithmic Learning Theory.

sources and construction of a statistical model for learning dependencies.

Ingrid Benti and Virginie Tumelaire worked on the automatic acquisition of lexical data from bilingual dictionaries, and then on the manual classification of sense indicators. This data is an extremely valuable resource for semantic disambiguation.

#### **6.4.5 Site Coordinator: Susan Armstrong**

Susan Armstrong was co-editor of the special issue of the Journal of Machine Learning Research, on machine learning approaches to shallow parsing, which was discussed above.

#### **6.4.6 Other Researchers at ISSCO**

Colleagues at ISSCO participated in the EU Transrouter project concerned with automating initial parts of the translation process. One of the tools developed, of direct relevance to the LCG project, was an automatic repetition detector within a text or set of texts. The output of this module serves as a good indicator of what kind of further processing is appropriate for a given text.

Pierrette Bouillon worked on a project for the automatic acquisition of semantic lexicons that resulted in a paper presented at CoNLL-2000 “Inductive Logic Programming for Corpus-Based Acquisition of Semantic Lexicons”.

Andrei Popescu-Belis has worked on the emergence of grammar amongst communities of artificial agents.

#### **6.4.7 Training Activities**

Together with coordinator Susan Armstrong and Pierrette Bouillon, Adelina Hild co-supervised Francois Legras, who had a two-month internship at ISSCO during the summer, 1999. Francois worked on developing an automated Word-Guesser which was designed to provide additional coverage for the words from the corpora not identified by means of the existing lexicons. The performance of the Word-Guesser was compared against that of a rule-based guesser with hand-crafted derivational rules for German and was found to have higher recall. Francois’s internship at ISSCO gave him an opportunity to work in the area of computational linguistics.

Two interns collaborated in three-month projects at ISSCO during the summer, 2000. Jerome Barois and Lionel Deglise, both coming from a computer science study in Brest, had the opportunity to learn about Natural Language Processing techniques. They worked on adapting and extending the core processing tools for web-based access. These tools provide the platform for linguistic annotation of texts as a basis for the automatic induction of the

categories and phrases.

During the period involved Alexander Clark gave several invited talks, which are itemized below. Franck Thollard supervised Master's students at the University of St. Etienne.

Geneva thus provided a total of nine months of training connected with LCG work.

#### **6.4.8 Collaboration**

The primary collaboration has been with the Tübingen site; Alexander Clark has been employed part-time at Tübingen and Franck Thollard has been employed for a period at ISSCO.

#### **6.4.9 Industrial Involvement**

ISSCO researchers participated in a collaborative project with Xerox and the World Intellectual Property Organisation (WIPO) on terminology extraction and document clustering. Methods are under development to filter out unlikely term candidate pairs by learning from a database of manually validated pairs. Work on document clustering using simple unsupervised algorithms has produced promising results for efficient routing of documents to translators. This work will continue beyond the end of the project.

### **6.5 Xerox Research Centre Europe, Grenoble, France**

The Xerox Research Centre Europe (XRCE) participated in the LCG project with one postdoc, Nicola Cancedda, for three years from the beginning of March 1999 to the end of February 2002.

#### **6.5.1 Postdoctoral Researcher: Nicola Cancedda**

Most of the scientific work on the project concerned the use of Explanation-Based Learning (EBL) for solving the problem of grammar specialisation: adapting a grammar to a specific domain by trading part of its coverage against a reduction in ambiguity in the most effective way. This work led to a method valid for a large class of unification-based grammar formalisms which gave excellent experimental results (Cancedda & Samuelsson 2000, Cancedda & Samuelsson 2001). The method was partially extended to deal with the problem of grammar adaptation, in which the coverage of the base grammar can also be increased to fit a new domain.

A second problem investigated in the context of the project consisted in the automatic annotation with chunking information of an aligned bilingual corpus relying on a chunker

for one of the two languages and a bilingual dictionary. Some experiments were conducted using the Xerox XIP parser for French and the bilingual French/Italian pair of the ELRA MLCC 1.0 parallel corpus.

In the last part of the project, the research work was focused on probabilistic models for syntactic structural disambiguation, with a special emphasis on attachment disambiguation problems. This led to the development of a stochastic attachment model and of an unsupervised procedure for estimating the corresponding parameters from a corpus. This work is documented in (Gaussier & Cancedda 2001*a*, Gaussier & Cancedda 2001*b*).

### **6.5.2 Local Project Coordinators: Christer Samuelsson and Eric Gaussier**

The work of the postdoc was supervised by Christer Samuelsson until July 2000, and by Eric Gaussier afterwards.

Besides supervising the work of the postdoc, Christer Samuelsson's research concerned mainly the study of a theoretically well-founded unifying framework for stochastic dependency grammars.

The work of Eric Gaussier concerns a wide spectrum of problems ranging from monolingual and cross-language information retrieval to document categorisation and clustering and to monolingual and multilingual thesaurus induction.

### **6.5.3 Other Researchers**

A large number of researchers conducted research related to the LCG project in the three years of its activity at XRCE. Besides Christer Samuelsson and Eric Gaussier, they include Andreas Eisele, Anette Frank, Boris Chidlovskii, and, to a lesser extent, David Hull, Greg Grefenstette, Jimi Shanahan, Jean-Michel Renders and Cyril Goutte.

### **6.5.4 Project Related Activities at XRCE**

The LCG postdoc was integrated in a research group of about 20 members, focusing on a number of different problems ranging from Computational Linguistics (rule-based and stochastic parsing, multilingual language generation, lexical semantics etc.) to Information Retrieval and to mathematical formalisms (e.g. Finite State Machines). A significant portion of them involved the use of Machine Learning techniques.

### **6.5.5 Training Activities**

Training activities involving the LCG postdoc include:

- ACAI'99 Summer School on Machine Learning in Chania, Greece;
- TeSTIA'2000 ELSNET Summer School on Text And Speech Triggered Information Access in Chios, Greece;
- ESSLLI'2001 Summer School in Logic, Language and Information in Helsinki, Finland;
- attendance to a number of tutorial sessions at international conferences and specialised workshops;
- attendance to internal tutorials and to a 5-days course on Finite-State tools for Computational Linguistics.

Nicola Cancedda, moreover, animated a reading group on Machine Learning in which a large number of scientific papers were analysed and discussed. In this context he also delivered a tutorial on Bayesian Networks.

Xerox's participation thus resulted in 2.5 months of training through Nicola Cancedda's work. In addition Xerox provided a week-long course on its finite-state tools to all LCG researchers in May, 2000 (see third annual report).

## 6.6 Groningen

This site has employed one Postdoc and two PhD students. This is an overview of their research and training activities as well as a summary of the work of the local coordinator and other related researchers.

### 6.6.1 Postdoctoral Researcher: Miles Osborne

Prior to starting this post, Miles was a Research Associate at the Cambridge University Computer Laboratory, working on the EU funded project *Sparkle*. There he built a grammar learner embedded in a large scale natural language processing system. This Minimal Description Length-based learner incrementally extended a large, manually written Definite Clause Grammar. The learner could be trained on raw text, or else text annotated with parsed corpora. Furthermore, the learner was capable of constraining the search space using a limited form of background knowledge.

Miles resigned his previous post on the 30th of September 1998 and started the LCG position on the 1st of October 1998. He accepted a post as lecturer at the University of Edinburgh as of 1 Jan 2000, but returned to the LCG project for two months in the summer of 2000.

His LCG related research activities are summarised as follows:

Adaptation of the Sparkle DCG learner for NP identification. This task was straightforward, given the fact that the learner acquired NP rules already. The main change was allowing it

to be trained on parsed corpora annotated with NP information.

Induction of DCGs modelled as random fields, following from the previous activity in that the Sparkle learner was used to acquire a *superset* of the rules to be learnt. However, unlike the Sparkle learner, which used a local, greedy search method, the LCG learner performs a global search for the optimum model. Apart from search issues, another divergence from the Sparkle learner is to model the feature-based parses produced by the superset of rules in terms of a random field (equivalently, a maximum entropy distribution). Parameters of the field are estimated using iterative scaling. The best *subset* of rules are then defined in a Bayesian manner as the subset that simultaneously minimises the description length of the model (rules and random field parameters) and the description length of the training set encoded in the random field model (random field likelihood probability). A superset of DCG rules (16k) were acquired, and existing iterative scaling code was adapted to deal with the large event spaces involved with the task. The goals of this research are (a) making random field estimation Bayesian (though a compression-based prior), (b) global optimisation of the learning task and (c) an increased understanding of the strengths and weaknesses of Maximum Entropy / Random Field Modelling. Results were presented in COLING 2000 (Osborne 2000). Based on Miles' random-fields work, Tony Mullen investigated whether random-field modelling would lead to competitive parse selection results (see also below for more on Mullen's work.)

In CoNLL-2000, Miles also presented his "Shallow Parsing as Part-of-Speech Tagging" paper. Also in conjunction with Rob Malouf, Miles Osborne has worked on the task of making maximum entropy efficient. The results were presented at CLIN 2000.

### 6.6.2 Predoctoral Researcher 1: Stasinou Konstantopoulos

Stasinou joined the project in October 1998 as a PhD student.

Stasinou has been using the Aleph Inductive Logic Programming (ILP) System developed in Oxford (Srinivasan 2001) to induce first order predicate logic descriptions of linguistic phenomena. He has experimented with both English syntax (Konstantopoulos 2000) and Dutch phonology (Konstantopoulos 2001, also to be published by WEB-SLS, <http://www.essex.ac.uk/web-sls/>).

Furthermore he has been developing and experimenting with a data-parallel version of Aleph, which is compiled with a version of Yap equipped with an MPI interface to parallelise the task of proving all the examples as part of the evaluation of each hypothesised clause. MPI (Message Passing Interface) is a specification for libraries that facilitate the communication between the nodes involved in a parallel computation.

Finally, he has been writing his PhD thesis, due for submission in early 2003.

Non-LCG academic activities included being one of the organisers of TABU-Dag 2000. TABU-Dag is an annual one-day conference on general linguistics, organised by the University of Groningen. The 2000 TABU-Dag took place on 16 June 2000 See

<http://www.let.rug.nl/tabu/> for more information.

He has also attended a short course on the MPI interface and has spent some time experimenting with and training on the 128-node Linux cluster available in the Computation Centre of the University's. He is also maintaining the HP-UX port of the YAP Prolog System for which Aleph is written and extending YAP with a Prolog interface to MPI libraries. He has also ported an independent attempt for data-parallel ILP (IndLog on a modified YAP compiler) to his interface, aiming at unifying the two efforts at a Prolog MPI interface.

Non academic activities within Alfa-Informatica included his acting library liason between the department and the faculty Library.

### **6.6.3 Predoctoral Researcher 2: Susanne Schoof**

Susanne Schoof joined the project in 2001 as a Ph.D. student.

Her research focusses on verbal complementation structures, in particular a syntactically motivated differentiation between superficially similar structures. Following the informal work of Bolkestein (late 1970s) Susanne elaborated a formalisation of her ideas within the framework of Head-Driven Phrase Structure Grammar (HPSG).

Susanne is also investigating the behavior of reflexive pronouns, particularly in connection with the two different structures for verbal complementation. In “raising” constructions, reflexives always occur as the accusative subject, never as the accusative object of the infinitival VP whilst in “control” constructions where the infinitival VP is transitive, the reflexive pronoun is always the object and never the subject of the embedded infinitive. Literature studies show that this phenomenon has not been described (or even remarked upon) earlier. She has worked out a preliminary formalisation which she will be refining during the following months.

Susanne has given poster presentations of her work on the two constructions (BCN poster day, Groningen, February 2001 and 2002 and the LATLING Conference, Amsterdam, June 2001) and a also talk at the Meeting of Dutch Classical Linguists (Katwijk, November 2001). Her submission to the HPSG conference (Seoul, August 2002) was also accepted.

### **6.6.4 Network Coordinator: John Nerbonne**

John Nerbonne has continued investigations into the application of unsupervised learning to the problem of dialect classification with (Heeringa, Nerbonne & Kleiweg 2001), (Nerbonne & Heeringa 2001) and Wilbert Heeringa and John Nerbonne *Dialect Areas and Dialect Continua*, to appear in *Language Variation and Change* 13, 2002.

He has also done work on the subjects of phonological learning using abduction (Tjong Kim Sang & Nerbonne 2000b) and neural networks (Stoianov & Nerbonne 2001).

He has also published on the subject of computer support for human language learning (Nerbonne 2002) and, finally, his review of Stefan Müller's (Müller 1999) HPSG treatment of German was published in the *Journal of Germanic Linguistics* 13, 2001.

In part due to the success of the LCG work, John Nerbonne was elected president of the international *Association for Computational Linguistics* in 2002.

### 6.6.5 Related Groningen researchers

Gertjan van Noord is the project manager of the NWO PIONIER project *Algorithms for Linguistic Processing*. ALP is developing a large-scale unification-based grammar of Dutch (Bouma, van Noord & Malouf 2001), where a particular focus has been the development of a stochastic disambiguation model. See <http://www.let.rug.nl/vannoord/alp/> for more information on ALP.

Rob Malouf joined the Groningen group in July, 1999 as part of a university project focused on computational modeling of behavior, a collaboration between Computer Science, Computational Linguistics, Biophysics and Philosophy. He has focused on applying machine learning to a part of the LCG grammar task, namely word order in adjectival phrases (Malouf 2000). In addition he has worked on efficient processing techniques (Malouf, Carroll & Copestake 200) and Maximum Entropy modeling (Malouf & Osborne 2001). Since July 2001, he has continued work on maximum entropy-based statistical natural language processing as part of the ALP project; results of this work will be presented during the Sixth Conference on Natural Language Learning (CoNLL-2002). As of January 2002, Malouf began a project aimed at developing improved parameter estimation techniques for complex models in the context of a fellowship from the Royal Dutch Academy of Arts and Sciences (KNAW).

Rob Koeling, who worked at LCG partner SRI Cambridge, successfully defended his PhD thesis on *Dialogue-Based Disambiguation* in January 2002 (Koeling 2002).

Tony Mullen successfully defended his PhD thesis on *Compositional Features and Feature Merging for Maximum Entropy-Based Parse Selection* in March 2002. His research is aimed at locating where overfitting is most likely, and most damaging. Results on parse selection have been presented in the NLP Pacific Rim Symposium (Mullen, Malouf & van Noord 2001).

Results on using random-field modelling for parse selection have been presented at CoNLL-2000 (Mullen & Osborne 2000) and also at the student ACL session in Hong Kong, 2000.

Gosse Bouma is a permanent staff member in Groningen who has attended LCG meetings and who has applied machine learning to the problem of grapheme-to-phoneme conversion (Bouma 2000).

### 6.6.6 Training activities

Miles Osborne prepared and conducted a course on Statistical Natural Language Processing here in Groningen for twenty students, providing twenty months of training.

Stasinos Konstantopoulos attended the 11th and 13th European Summer School in Logic, Language and Information (Utrecht, 9-20 August 1999 and Helsinki, August 2001, resp. See also <http://ess11i.let.uu.nl>.) He has also attended various talks on machine learning (e.g. the one of Ray Mooney, University of Houston, USA in Tilburg) or subjects of general interest (e.g. the weekly Taalkundig Colloquium in Groningen). He is also assisting Gosse Bouma by giving tutorials for the NLP I course during the spring trimester of 2000, 2001 and 2002 (see <http://www.let.rug.nl/gosse/nlp1/> for more details on the course), providing a total of 21 person months of training (figuring his contribution as 20% of the total, which was 35 months/year for three years).

He has, finally, been given the chance to make a short visit to the University of York in the summer of 2000, where he met James Cussens, a researcher who is very active in ILP in general and the development of Aleph in particular.

Susanne Schoof has attended the BCN Introductory course, offered to new Ph.D. students, and she assisted in an introductory course in information science.

### 6.6.7 Industrial Involvement

Combining inter-LCG collaboration with industrial involvement, Rob Koeling (SRI Cambridge) worked together with Gosse Bouma (Alfa-Informatica), Tanja Gaustad (Alfa-Informatica, ALP project) and John Nerbonne on an email classification project<sup>2</sup> contracted by the Groningen-based company, BSC<sup>3</sup>. Rob Koeling's involvement in the email classification project lasted from September 2000 until February 2001, and for the purposes of the project he was visiting Groningen in January and February 2001.

### 6.6.8 Collaboration

John Nerbonne, Hervé Déjean (Tübingen) and Erik Tjong Kim Sang (Antwerp) have jointly given the *Learning Computational Grammars* paper in CoNLL-2001 in Toulouse (Nerbonne et al. 2001).

Nathan Vaillette (Tübingen) delivered a talk in April 2001 regarding an interpretation of Monadic Second Order Logic under which it is equivalent to regular expressions. The talk described an implementation of MSOL under this interpretation using Gertjan van Noord's

---

<sup>2</sup><http://vili.let.rug.nl/zonnet/>

<sup>3</sup><http://www.bsc.nl/>

FSA Utilities<sup>4</sup>. The talk was given in the context of the weekly colloquium of CLCG<sup>5</sup>.

Frank Thollard (Tübingen) also visited Groningen in September 2001 to deliver a talk *On the understanding of algorithms Alergia, MDI and DDSM*, also for the CLCG colloquium. Frank's talk immediately preceded a reading group organised within Alfa-Informatica on FSA induction in general and the *Alergia* algorithm in particular.

## 6.7 Tübingen

The Tübingen site has had, over the course of the project, five different young researchers. Two of these five, Hervé Déjean and Franck Thollard have made the most direct contribution to the goals of the project. Yuval Krymolowski from Israel was invited for a shorter period to consult with young researchers on the topic of memory-based learning for the TMR-LCG task. Alexander Clark was only part time in Tübingen since he worked primarily at the Geneva site. In Tübingen he collaborated very closely with Franck Thollard on using grammatical induction for the chunking task. The final young researcher in Tübingen was Wouter Jansen from LCG partner Groningen, who applied ideas from the LCG chunking task to the domain of phonetics.

### 6.7.1 Postdoctoral Researcher 1: Hervé Déjean

Dr. Déjean's involvements with LCG started at Tübingen 1st February 1999 and ended 31 March 2001 (for a position at Xerox, Grenoble, LCG's industrial partner). His activities during his stay in Tuebingen can be summarized as follows:

The main activity during these two years was the conception and development of a top-down rule-inducing (and theory refining) system, ALLiS. Top-down induction is a generate-then-test approach: a set of potential rules is generated according to some patterns, and each of them is tested against the training data. Those being accurate enough are kept, the others deleted. The major difference with other traditional inductive systems concerns the termination criteria: whenever a rule is learned, ALLiS tests whether exceptions to this rule can be learned as well.

While most algorithms learn a set of (ordered) rules, ALLiS learns a structured set of rules. With each rule is explicitly associated a set of exceptions. As different results show, the accuracy of the structure (rule + exceptions) is generally greater than the accuracy of the single rule itself. Online demo of the chunk parser can be found at the URL: <http://www.sfb441.uni-tuebingen.de/~dejean/lcg/chunker.html>. ALLiS was successfully applied to the CoNLL-2000 shared task, namely chunking. Results were presented during CoNLL-2000 in Lisbon. They show that ALLiS is competitive with other systems, and it gets the best score of all the symbolic learning systems.

---

<sup>4</sup><http://odur.let.rug.nl/vannoord/Fsa/>

<sup>5</sup>See <http://www.let.rug.nl/clcg/> about CLCG and the colloquium

ALLiS is now successfully used at Xerox in several projects, where entities recognition is required.

In collaboration with Erik Tjong Kim Sang (Antwerp University), he prepared the CoNLL-2001 shared task: Clausing, namely a segmentation into clauses. ALLiS was also applied on this new data.

Dr. Déjean attend several international conferences (EACL'99, CoNLL'99, LREC'00, COLING'00, CoNLL'01), where he presented his work. He wrote 5 articles presented at international conferences, and one journal article.

### **6.7.2 Postdoctoral Researcher 2: Franck Thollard**

Franck Thollard joined the project in September 2001. He first joined the Tübingen site and then later went in Geneva in order to work with Alexander Clark. Franck Thollard worked in two directions : on the one hand he adapted his grammatical inference technique to a shared task on Noun Phrase Chunking. This task was provided by some TMR researchers (mainly Erik Tjong Kim Sang). On the other hand, he worked on improving grammatical inference algorithms.

By the end of the time spent in Tübingen, his system had reached the level of 87% accuracy. These results led to an article at the French machine learning conference (Thollard 2001*b*) and participation in the presentation of an invited paper at the CoNLL-2001 conference (Nerbonne et al. 2001).

Franck Thollard also worked on improving grammatical inference algorithms. This led to an article at the International Colloquium on Machine Learning (Thollard 2001*a*).

During a meeting in Tübingen, Franck Thollard and Alexander Clark found great opportunities for collaboration. In order to further this collaboration, Alexander Clark came to work part time in Tübingen, and Franck Thollard spent a large portion of his final year in Geneva.

Before beginning this collaboration with Geneva, Franck Thollard was invited by Gertjan Van Noord (University of Groningen) and presented a lecture on grammatical inference. Some collaborations started there and are still being pursued.

Moreover, Franck Thollard started a collaboration with the EURISE Team (Saint Etienne, France). This led to a paper at the European Conference on Machine Learning (2002). Franck Thollard has also collaborated with the Master Thesis of the university of Saint Etienne and supervised Toufik Boudellal who collaborates with Sonia Halimi (University of Geneva) and Alexander Clark.

### 6.7.3 Predoctoral Researcher 1: Alexander Clark

Dr. Clark worked part time in Tübingen in collaboration with Franck Thollard. Together they worked out details of applying grammatical inference to chunking and other related tasks. The details of this collaboration, along with resulting publications, are presented in the Geneva section of this report.

### 6.7.4 Predoctoral Researcher 2: Yuval Krymolowski

Yuval Krymolowski was a predoctoral researcher in Tübingen between May and August 2001. He is a PhD student at the department of computer science in Bar-Ilan university, Israel. His supervisor is Ido Dagan. As a thesis project, Yuval is developing a memory based system for compositional partial parsing (Krymolowski & Dagan 2000, Krymolowski & Dagan 2001). The system, called MBSL for “Memory-Based Sequence Learning” is designed for learning phrases with internal structure such as VPs which contains NPs, or NPs composed of internal phrases. An earlier version of MBSL (Argamon, Dagan & Krymolowski 1999) was used for chunking. Both learning tasks, of chunks and compositional relations, are also part of the LCG project, and MBSL was one of the systems participating in the CoNLL-2000 shared task. During his stay in Tübingen, he worked primarily on extending MBSL for handling dependency relations.

Yuval Krymolowski furthermore worked on developing a view of statistical NLP systems as samples of statistical variables. As such, their performance is itself a statistical variable, and one can study its distribution and correlation properties. Such study is related to the issue of system transferability. When using a system trained, e.g., on economical data such as WSJ, it is important to estimate the spread of its performance over similar kinds of data. Preliminary results of this study were presented in ACL-2001 workshop on Evaluation for Language and Dialogue Systems (Krymolowski 2001).

While at Tübingen he worked most closely with Sandra Kübler, a PhD student who is working on her own memory-based parsing system. This collaboration was of mutual benefit for their PhD research. Yuval also presented the MSBL system and gave a seminar presentation on related memory-based approaches in a postgraduate course on statistical methods for parsing.

Yuval maintains ongoing contacts with colleagues in Tilburg regarding shallow parsing techniques and methodology. In early August he presented and discussed his evaluation paper with ILK group members, including Walter Daelemans and Erik Tjong Kim Sang from Antwerp.

### **6.7.5 Predoctoral Researcher 3: Wouter Jansen from Groningen**

Wouter Jansen was the final young researcher to be employed at Tübingen. His work applies core chunking ideas such as boundary markers for segmentation to the domain of phonetic analysis. His work thus exemplifies a possible direction for future work developed from the LCG work on chunking.

Wouter Jansen was employed at Tübingen between Sept 2001 and March 2002. He is currently a PhD student in the departments of Linguistics and Humanities Computing at the University of Groningen. His main research interests have been grammatical boundary phenomena in phonology and phonetics, such as the distribution of secondary stresses at the compound and phrase levels, and segmental assimilation rules. Knowledge of boundary phenomena has been shown to be useful in chunking the speech signal into grammatical constituents during speech recognition (Cutler 1996).

During Wouter Jansen's employment in Tübingen, he had the opportunity to learn about finite-state implementations of Optimality Theory (Prince & Smolensky 1993) from Dale Gerdemann, which directly benefited his work on optimality-theoretic models of English phrase and (nominal) compound stress. During his time in Tübingen, he also collaborated with Katja Jasinskaja, a PhD student in the Department of Linguistics, on a psycholinguistic experiment concerning the effects of different acoustic reflexes of phonological phrasing on semantic interpretation.

### **6.7.6 Site Coordinator: Dale Gerdemann**

Dale Gerdemann has coordinated LCG research at Tübingen. His expertise is primarily in the area of finite state methods in natural language processing, a topic which has been of central importance in the Tübingen LCG research. Both the work of Hervé Déjean and Franck Thollard explicitly used finite state models. Thus a considerable amount of collaboration and exchange of ideas was possible.

In 2000/2001 he organized a seminar along with Hervé Déjean and Franck Thollard on Machine Learning and Finite State Methods in NLP. And in 2001, he co-taught (along with Prof. Erhard Hinrichs) a seminar in statistical methods in parsing. Franck Thollard and Yuval Krymolowski participated in this seminar along with local students, who benefited from learning about the LCG project. Both of these courses were primarily for the benefit of the TMR researchers.

Dale Gerdemann was also actively involved in training Wouter Jansen in finite-state methods and Optimality Theory, which was of important benefit for Wouter's PhD thesis. Wouter Jansen's application of Optimality Theory to phonetics was very useful for Dr. Gerdemann's ongoing research in computational implementations of this theory. Wouter Jansen's research activities in Tübingen served to strengthen collaborative ties with his home university in Groningen.

Dale Gerdemann further collaborated with Gertjan van Noord of Groningen University on finite state methods. Their collaboration has recently resulted in a publication in the journal *Grammars* (van Noord & Gerdemann n.d.).

### 6.7.7 Related Tübingen Researchers

Prof. Dr. Erhard Hinrichs and Sandra Kübler, M.A. (staff member) have both been deeply involved with LCG and the LCG researchers over the past 4 years. Together they have worked on similarity based approaches to chunking, which are similar to the memory based learning approach of the Antwerp site and of the LCG predoc, Yuval Krymolowski. Their work is recorded in several publications (Hinrichs, Kübler, Müller & Ule 2002, Kübler & Hinrichs 2001*a*, Kübler & Hinrichs 2001*b*, Kübler 2001). Recently they organized a workshop on "Machine Learning Approaches in Computational Linguistics" held at the ESSLLI 2002 summer school in Trento.

Tylman Ule (staff member) participated in the seminar on statistical parsing methods along with Franck Thollard and Yuval Krymolowski. He applied parsing techniques acquired in this course and elsewhere for the robust finite-state chunk parser used in the DEREKO (Deutsches ReferenzKorpus) project. He benefited greatly from many discussions with both Frank Thollard and Hervé Déjean on topics related to XML text annotation, Bagging/Boosting and C++ implementations.

Frank Müller (staff member) has worked closely with Tylman Ule on parsing methods and text annotation. He shared many discussions with TMR postdocs Franck Thollard and Hervé Déjean, and benefited from attending the 2000 Xerox Grenoble tutorial on XFST and related software.

Klaus Hörmann is an MA student who participated in seminars on finite-state methods and statistical parsing along with Franck Thollard, Yuval Krymolowski and Hervé Déjean. His masters thesis, which he recently completed, is on constraint grammar for morphological disambiguation. The approach he used involves learning by successive refinement, an approach similar to that of Hervé Déjean.

Nathan Vaillette is a PhD student from Ohio State University who visited Tübingen during the third year of the LCG project. As reported in the third year report, he participated in a seminar in finite state methods for natural language processing, which was led by Dale Gerdemann and Franck Thollard. As a project for this course, he implemented extensions to the finite state calculus to handle monadic second order logic, which he reported on in an invited talk in Groningen. Since then (as reported in the fourth year report), he has presented at the international Finite State Methods in Natural Language Processing workshop at ESSLI in Helsinki (Vaillette 2001).

### 6.7.8 Training Activities

Over the course of the TMR-LCG project, Tübingen has had five different young researchers, who have received and provided training totalling nine person-months. Over the course of the project, two courses were taught which were primarily for the benefit of the TMR researchers. In the winter semester of 2000, a seminar course held on machine learning and finite state methods in NLP. Franck Thollard and Hervé Déjean were both involved in this course. And in the summer semester of 2001, a course was held on statistical methods and parsing, which included participation of Franck Thollard, Hervé Déjean and Yuval Krymolowski. Franck Thollard and Hervé Déjean additionally gave talks in other courses, both in the Linguistics department as well as the Computer Science department. Additionally, Hervé Déjean was involved in a machine learning reading group for a full semester.

Given the inexperience of the young researchers, it was necessary to provide individual training in research methods along with writing and presenting of results. This training contributed to the success of the Tübingen young researchers in publishing their results. Wouter Jansen, in particular, started very late in the project and received primarily individual training in finite state methods and applications to computational phonology. Wouter Jansen in turn provided training in experimental design to a Ph.D. student in Tübingen and Hervé Déjean was involved with supervising a student software practicum of Klaus Hörmann.

Alexander Clark is a special case since he was only part time in Tübingen. His training in machine learning methods came primarily through his collaboration with Franck Thollard, who spent 6 months at the Geneva site before returning to Tübingen for the final two months of the project.

The Tübingen LCG researchers and related young researchers have also received training outside of Tübingen. In 2000 Hervé Déjean and Frank Müller attended the four day Xerox Grenoble tutorial on XFST and related NLP software. Also in 2000 Franck Thollard and Hervé Déjean attended a Tutorial (EAIA'00, Lisbon, September 2000) on the topic: Information Exploration and Learning Text. In 1999, Sandra Kübler visited the Antwerp site and was trained in the use of TiMBL. The LCG researchers also benefited from project reports and other lectures at regular LCG meetings and from visits of LCG researchers from other sites.

### 6.7.9 Industrial Involvement

The most important industrial involvement for Tübingen has been with LCG partner Xerox. Through this involvement, the young researcher Hervé Déjean was successful in attaining a position at the company. Also as a result of discussions with Lauri Karttunen at the spring 2000 meeting in Grenoble, Dale Gerdemann received an invitation to give a talk in COLING on work done in collaboration with Gertjan van Noord of LCG partner Groningen, with whom he developed implementations of finite-state optimality theory (Gerdemann & van

Noord 2000).

### 6.7.10 Collaborations

Numerous collaborations and cross-fertilizations of ideas have developed over the course of the project. For Tübingen, the most important collaboration has been between Franck Thollard and Alexander Clark. This collaboration resulted in Dr. Clark working partly in Tübingen and Franck Thollard partly in Geneva. Details of this collaboration are given in both the Tübingen and the Geneva sections of this report.

An ongoing collaboration between Dale Gerdemann (site coordinator, Tübingen) and Gertjan van Noord (Groningen) has resulted in a number of publications and conference presentations on finite state methods in natural language processing. Partly as a result of this collaboration, the LCG postdoc Franck Thollard and related young researcher Nathan Vaillette were both invited to give presentations and tutorials in Groningen.

## 7 Future Prospects

LCG was clearly correct in its premise that machine learning techniques would become a fundamental technology in natural language processing. This is evident in sessions of professional meetings, in the submissions to journals, and in the demand for professionals with this expertise. Applications of machine learning to natural language processing have emerged as a major theme at contemporary scientific meetings.

Several emerging questions would be suitable for future large-scale collaboration, including the examination of linguistic problems more challenging than shallow syntax; the inclusion of techniques which have only recently emerged, e.g., kernel techniques (support vector machines); and the more extensive use of elaborate linguistic assumptions in tandem with learning algorithms.

A more novel approach might focus on the degree to which unsupervised techniques may be exploited. This question is motivated scientifically by the fact of human language acquisition, which proceeds in spite of children obtaining little feedback, and ignoring what little they do get. It is likewise motivated practically by the huge amount of data available for unsupervised training, and the scarcity and expense of obtaining the annotations required for supervised learning.

It would also be worthwhile to investigate the application of learning techniques to natural data, i.e., the data available to a child learning language. Child language acquisition is already a popular field in which computer simulation is applied, but machine learning techniques are curiously underrepresented.

## References

- Akhtar, S. & R. G. Reilly (2001), Automating XML mark-up, *in* 'Proceedings of Conference on Humanities Computing, New York University, June'.
- Argamon, S., I. Dagan & Y. Krymolowski (1999), 'A memory-based approach to learning shallow natural language patterns', *Journal of Experimental and Theoretical AI* 11, 369–390. CMP-LG/9806011.
- Belz, Anja (2000), Multi-syllable phonotactic modelling, *in* 'Proceedings of SIGPHON 2000: Finite-State Phonology', pp. 46–56.
- Belz, Anja (2001a), Learning local structural context grammars for different parsing tasks by partition-tree search, Technical Report SRI-TT-02-5, SRI Cambridge.
- Belz, Anja (2001b), Optimisation of corpus-derived probabilistic grammars, *in* 'Proceedings of Corpus Linguistics 2001', Lancaster, UK, pp. 46–57.
- Belz, Anja (2002), Learning broad-coverage grammars for different parsing tasks from parsed corpora: Report on research for leg project 'learning computational grammars', Technical Report SRI-TT-02-31, SRI Cambridge.
- Bouma, Gosse (2000), A finite-state and data-oriented method for grapheme to phoneme conversion, *in* 'Proceedings of the first conference of the North-American Chapter of the Association for Computational Linguistics', Association for Computational Linguistics, Somerset, NJ, pp. 303–310.
- Bouma, Gosse, Gertjan van Noord & Robert Malouf (2001), Alpino: Wide-coverage computational analysis of dutch, *in Proceedings of Computational Linguistics in the Netherlands 2000 (Proceedings of Computational Linguistics in the Netherlands 2000 2001)*.
- Callan, R.E. & D. Palmer-Brown (1997), '(S)RAAM: An analytical technique for fast and reliable derivation of connectionist symbol structure representations', *Connection Science* 9(2), 139–159.
- Cancedda, Nicola & Christer Samuelsson (2000), Experiments with corpus-based lfg specialization, *in* 'Proceedings of the NAACL-ANLP 2000 Conference', Seattle, WA.
- Cancedda, Nicola & Christer Samuelsson (2001), Corpus-based grammar specialisation, *in* 'Proceedings of the Fourth Conference on Computational Natural Language Learning (CoNLL 2001)', Lisbon, Portugal.
- Cutler, Anne (1996), Prosody and the word boundary problem, *in* J.Morgan & K.Demuth, eds, 'Signal to syntax: Bootstrapping from speech to grammar in early acquisition', Erlbaum, Hillsdale, NJ, pp. 87–99.
- Daelemans, Walter, Sabine Buchholz & Jorn Veenstra (1999), Memory-based shallow parsing, *in* 'Proceedings of CoNLL-99', Bergen, Norway, pp. 53–60.
- Gaussier, Eric & Nicola Cancedda (2001a), Probabilistic models for pp-attachment resolution and np analysis, *in* W.Daelemans & R.Zajac, eds, 'Proceedings of the 5th Workshop on Computational Natural Language Learning (CoNLL-2001)', The Association for Computational Linguistics, Toulouse, France.
- Gaussier, Eric & Nicola Cancedda (2001b), Probabilistic models for terminology extraction and knowledge structuring from documents, *in* 'Proceedings of the 2001 IEEE Workshop on Natural Language Processing for Knowledge Engineering (NLPKE 2001)', Tucson, Arizona.
- Gerdemann, Dale & Gertjan van Noord (2000), Approximation and exactness in finite state optimality theory, *in* 'Proceedings of Sigphon Workshop on Finite State Phonology', Luxembourg. Invited paper.

- Hammerton, J. (2001), Clause identification with Long Short-Term Memory, *in* ‘Proceedings of the CoNLL 2001 workshop, ACL 2001, Toulouse, France’.
- Hammerton, J. A. (1999), Holistic Symbol Processing, *in* D.Bridge, R.Byrne, B.O’Sullivan, S.Prestwich & H.Sorensen, eds, ‘Pre-proceedings of the Tenth Irish Conference on Artificial Intelligence and Computer Science, Sept 10-13, University College Cork’, Dept. of Computer Science, University College Cork, Cork, Ireland.
- Hammerton, James & Erik F. Tjong Kim Sang (2001), Combining a self-organising map with memory-based learning, *in* ‘Proceedings of CoNLL-2001’, Toulouse, France, pp. 9–14.
- Heeringa, Wilbert, John Nerbonne & Peter Kleiweg (2001), Validating dialect comparison methods, *in* W.Gaul & G.Ritter, eds, ‘Proceedings of the 24th Annual Conference of the Gesellschaft für  $\frac{1}{2}$ r Klassifikation’, Classification, Automation, and New Media.
- Heffernan, P. & J. Hammerton (12–15 December, 2000), Holistic Unification and the Bi-coding RAAM, *in* ‘Intelligent Systems and Architectures (ISA 2000)’, University of Wollongong, Australia.
- Hinrichs, Erhard W., Sandra Kübler, Frank H. Müller & Tylman Ule (2002), A hybrid architecture for robust parsing of german, *in* ‘Proceedings of LREC’, Las Palmas, Gran Canaria.
- Hochreiter, S. & J. Schmidhuber (1997), ‘Long short-term memory’, *Neural Computation* **9**, 1735–1780.
- Hoste, Véronique, Walter Daelemans, Erik Tjong Kim Sang & Steven Gillis (2000), Meta-learning for phonemic annotation of corpora, *in* ‘Machine Learning: Proceedings of the Seventeenth International Conference (ICML-2000)’, Stanford CA, USA, Morgan Kaufmann, pp. 375–382.
- James, D. L. & R. Miikkulainen (1995), *SARDNET: A Self-Organizing Feature Map for Sequences*, MIT Press, Cambridge, MA, pp. 577–584.
- Kechadi, M. T. & Reilly R.G. (2000), ‘Parallel implementation of an unconstrained optimisation learning algorithm’, *Journal of Parallel and Distributed Systems* **10**(4).
- Koeling, Rob (2000a), Chunking with maximum entropy models, *in* ‘Proceedings of CoNLL-2000’, Lisbon, Portugal.
- Koeling, Rob (2000b), Using dialogue information for parsing in a spoken dialogue system, *in* ‘Proceedings of Gotalog 2000: Fourth workshop on the semantics and pragmatics of dialogue’.
- Koeling, Rob (2002), Dialogue-Based Disambiguation: Using Dialogue Status to Improve Speech Understanding, PhD thesis, Maths and Sciences, Rijksuniversiteit Groningen.
- Konstantopoulos, Stasinos Th. (2000), NP chunking using ILP, *in* P.Monachesi, ed., ‘Proceedings of Computational Linguistics in the Netherlands 1999’, Utrecht Institute of Linguistics OTS, Utrecht, pp. 109–116.  
**URL:** <ftp://ftp.let.rug.nl/pub/konstant/Docs/clin1999.ps.bz2>
- Konstantopoulos, Stasinos Th. (2001), Learning phonotactics using ILP, *in* K.Striegnitz, ed., ‘Proc. of the Sixth ESSLI Student Session’, Helsinki, pp. 148–158.  
**URL:** <ftp://ftp.let.rug.nl/pub/konstant/Docs/esslli01.ps.bz2>
- Krymolowski, Y. (2001), Using the distribution of performance for studying statistical nlp systems and corpora, *in* ‘Proceedings of ACL Workshop on Evaluation Methodologies for Language and Dialogue Systems’, Toulouse, France, pp. 52–59.
- Krymolowski, Y. & I. Dagan (2000), Incorporating compositional evidence in memory-based partial parsing, *in* ‘ACL00’, Hong Kong, pp. 45–52.
- Krymolowski, Y. & I. Dagan (2001), Compositional memory-based partial parsing, *in* R. S.R. Bod & K.Sima’an, eds, ‘Data-Oriented Parsing’, CSLI Publications, chapter II.7. invited, *in* print.

- Kübler, Sandra (2001), Braucht nominalphrasenerkennung linguistisches wissen?, in 'Proceedings der GLDV-Frühjahrstagung', Gießen.
- Kübler, Sandra & Erhard W. Hinrichs (2001a), From chunks to function-argument structure: A similarity-based approach, in 'Proceedings of ACL-EACL'.
- Kübler, Sandra & Erhard W. Hinrichs (2001b), Tüsl: A similarity-based chunk parser for robust syntactic processing, in 'Proceedings of HLT', San Diego, Cal.
- Malouf, Robert (2000), The order of prenominal adjectives in natural language generation, in 'Proceedings of 38th Annual Meeting of the Association for Computational Linguistics', Hong Kong, pp. 85–92.
- Malouf, Robert, John Carroll & Ann Copestake (200), 'Efficient feature structure operations without compilation', *Natural Language Engineering* 6(1), 29–46.
- Malouf, Robert & Miles Osborne (2001), A toolkit for robust and efficient maximum entropy language modeling., in *Proceedings of Computational Linguistics in the Netherlands 2000* (*Proceedings of Computational Linguistics in the Netherlands 2000* 2001).
- Mayberry, M. & R. Miikkulainen (1998), SARDSRN: A neural-network shift-reduce parser, Technical Report AI98-275, Department of Computer Science, University of Texas at Austin, Texas, US.
- Mullen, Tony & Miles Osborne (2000), Overfitting avoidance for stochastic modeling of attribute-value grammars, in J.Cussens & S.Džeroski, eds, 'Proceedings of CoNNL2000 and LLL2000', Lecture Notes in Artificial Intelligence, Springer, Berlin.
- Mullen, Tony, Robert Malouf & Gertjan van Noord (2001), Statistical parsing of Dutch using maximum entropy models with feature merging, in 'Proceedings of the 6th Natural Language Processing Pacific Rim Symposium', Tokyo.
- Müller, Stefan (1999), *Deutsche Syntax deklarativ: Head-Driven Phrase Structure Grammar für das Deutsche*, number 394 in 'Linguistische Arbeiten', Max Niemeyer Verlag, Tübingen.
- Nenova, N. & R. G. Reilly (2001), A taxonomy of discourse particles in spoken language, in 'Proceedings of the Workshop on Discourse Particles in Speech, Brussels, Belgium'.
- Nerbonne, John (2002), Computer-assisted language learning and natural language processing, in R.Mitkov, ed., 'Handbook of Computational Linguistics', Oxford University Press.
- Nerbonne, John, Anja Belz, Nicola Cancedda, Hervé Déjean, James Hammerton, Rob Koeling, Stasinou Konstantopoulos, Miles Osborne, Franck Thollard & Erik Tjong Kim Sang (2001), Learning computational grammars, in W.Daelemans & R.Zajac, eds, 'Proceedings of CoNLL-2001', Toulouse, France, pp. 97–104.
- Nerbonne, John & Wilbert Heeringa (2001), 'Computational comparison and classification of dialects', *Dialectologia et Geolinguistica* 9.
- Osborne, M. (2000), Estimation of stochastic attribute-value grammars using an informative sample, in 'Proceedings of COLING 2000', Saarbrücken, pp. xxx–yyy.
- Prince, Alan & Paul Smolensky (1993), Optimality Theory: Constraint interaction in generative grammar, Technical Report 2, Center for Cognitive Science, Rutgers University.
- Proceedings of Computational Linguistics in the Netherlands 2000* (2001).
- Reilly, R. G. & D. Mackey (2001), Cortical software re-use: A theory of cognitive development, in 'Fifth International Conference on Cognitive Science and, neural systems'.
- Reilly, R.G. (1999), 'A case study of transient dyslexia.', *Brain and Language* 70, 336–346.
- Reilly, R.G. (2000), Evolution of symbolisation: Signposts to a bridge between connectionist and hybrid symbolic systems, in S.Wermter & R.Sun, eds, 'Hybrid Neural Systems', Springer-Verlag.

- Reilly, R.G. ((in press)), ‘The relationship between object manipulation and language development in broca’s area: A connectionist simulation of greenfield’s hypothesis.’, *Behavioral and Brain Sciences*.
- Srinivasan, Ashwin (2001), *Aleph*, <http://www.comlab.ox.ac.uk/oucl/research/areas/machlearn/Aleph/>.
- Stoianov, Ivilin & John Nerbonne (2001), Learning lexical phonotactics with simple recurrent networks. submitted to *Computer Speech and Language*.
- Thollard, Franck (2001a), Improving probabilistic grammatical inference core algorithms with post-processing techniques, in ‘Eighth Intl. Conf. on Machine Learning’, Morgan Kaufmann, Williams, pp. 561–568.
- Thollard, Franck (2001b), Inférence grammaticale probabiliste et détection de groupes nominaux : résultats préliminaires, in PUG, ed., ‘Conférence d’Apprentissage (CAp 2001)’, Plate-forme AFIA, Gilles Bissons, Grenoble, pp. 227–242.
- Tjong Kim Sang, Erik F. (2000a), Noun phrase recognition by system combination, in ‘Proceedings of the ANLP-NAACL 2000’, Seattle, Washington, USA. Morgan Kaufman Publishers, pp. 50–55.
- Tjong Kim Sang, Erik F. (2000b), Text chunking by system combination, in ‘Proceedings of CoNLL-2000 and LLL-2000’, Lisbon, Portugal, pp. 151–153.
- Tjong Kim Sang, Erik F. (2001a), Memory-based clause identification, in ‘Proceedings of CoNLL-2001’, Toulouse, France, pp. 67–69.
- Tjong Kim Sang, Erik F. (2001b), Transforming a chunker to a parser, in ‘Computational Linguistics in the Netherlands 2000’, Tilburg, The Netherlands, pp. 177–188.
- Tjong Kim Sang, Erik F. (2002), ‘Memory-based shallow parsing’, *Journal of Machine Learning Research* **2**(Mar), 559–594.
- Tjong Kim Sang, Erik F. & Hervé Déjean (2001), Introduction to the conll-2001 shared task: Clause identification, in ‘Proceedings of CoNLL-2001’, Toulouse, France, pp. 53–57.
- Tjong Kim Sang, Erik F. & John Nerbonne (1999), Learning simple phonotactics, in ‘Proceedings of the Workshop on Neural, Symbolic, and Reinforcement Methods for Sequence Processing’, ML2 workshop at IJCAI’99, Stockholm, Sweden, pp. 41–46.
- Tjong Kim Sang, Erik F. & John Nerbonne (2000a), Learning the logic of simple phonotactics, in ‘Learning Language in Logic’, Vol. 1925 of *Lecture Notes in Computer Science*, Springer Verlag, pp. 110–124.
- Tjong Kim Sang, Erik F. & Jorn Veenstra (1999), Representing text chunks, in ‘Proceedings of EACL’99’, Bergen, Norway, pp. 173–179.
- Tjong Kim Sang, Erik F. & Sabine Buchholz (2000), Introduction to the conll-2000 shared task: Chunking, in ‘Proceedings of CoNLL-2000 and LLL-2000’, Lisbon, Portugal, pp. 127–132.
- Tjong Kim Sang, Erik F., Walter Daelemans, Hervé Déjean, Rob Koeling, Yuval Krymolowski, Vasin Punyakanok & Dan Roth (2000a), Applying system combination to base noun phrase identification, in ‘Proceedings of COLING 2000’, Saarbruecken, Germany, pp. 857–863.
- Tjong Kim Sang, Erik F., Walter Daelemans, Hervé Déjean, Rob Koeling, Yuval Krymolowski, Vasin Punyakanok & Dan Roth (2000b), Applying system combination to base noun phrase identification, in ‘Proceedings of the Coling 2000’, Association for Computational Linguistics.
- Tjong Kim Sang, Erik & John Nerbonne (2000b), Learning the logic of simple phonotactics, in J.Cussens & S.Džeroski, eds, ‘Learning Language in Logic’, Vol. 1925 of *Lecture Notes in Artificial Intelligence*, Springer Verlag.

- Vaillette, Nathan (2001), Logical specification of transducers for nlp, *in* 'Proceedings on FSMNLP 2001', University of Helsinki.
- van Noord, Gertjan & Dale Gerdemann (n.d.), 'Finite state transducers with predicates and identity', *Grammars* 4(3), 263–286.