Learning Computational Grammars
TMR Project Nr. ERBFMRXCT980237
3rd Annual Report

John Nerbonne

September 20, 2001

Summary

This document describes the third year’s progress of the TMR Project Learning Computational Grammars (LCG). In brief, LCG continues with full complement of postdocs and predocs, and work in areas as diverse as Maximum Entropy, Instance-based Learning, Neural Networks, Explanation-Based Learning, Theory Refinement, Inductive Logic Programming, and Genetic Algorithms. In keeping with the original project proposal, most sites continue to targeting their various learning technologies on the task of learning noun phrases in free text. The industrial partner, Xerox, is exploring an application, and Geneva has switched focus from linguistic and psycholinguistic accounts of learning to unsupervised machine learning techniques.

A highlight of the third year has been that the network opened participation in one of its “shared tasks” to the international scientific community at the Sept. 2000 CoNLL meeting, resulting in the publication of twelve papers describing various approaches. LCG created a task description and an attendant training and testing set, which was the focus of the public meeting, held in conjunction with the Special Interest Group in Natural Language Learning (CoNLL), the Language Learning in Logic conference (LLL), and the International Conference on Grammar Induction (ICG1). Project members accounted for four of the world’s seven best results in the so-called “text-chunking” task.
Contents

1 Introduction 5

2 Training 5

3 Industrial Involvement 6

4 Collaborations 6

5 Meetings 7

5.1 Grenoble, May 2000 ................................................... 7
5.2 Tübingen, February 2001 ............................................. 8
5.3 Lisbon, September 2000 .............................................. 8
5.4 ICGI-2000, Lisbon, September 2000 ............................. 10
5.5 CoNLL, Lisbon, September 2000 .................................. 10
5.6 LLL, Lisbon, September 2000 ....................................... 10

6 Site Reports 11

6.1 Antwerp ................................................................. 11

6.1.1 Erik Tjong Kim Sang, Postdoc ............................... 11
6.1.2 Walter Daelemans, Coordinator ............................. 12
6.1.3 Training Activities .............................................. 12

6.2 SRI International, Cambridge, UK ............................... 13

6.2.1 Summary of LCG Project Activities at SRI ............... 13
6.2.2 Anja Belz, Postdoctoral Researcher......................... 14
6.2.3 Rob Koeling, Doctoral Researcher ......................... 14
6.2.4 David Milward, Coordinator ........................................... 15
6.2.5 Involvement of other Researchers ..................................... 15
6.2.6 New Publications ....................................................... 15
6.2.7 Training Activities ..................................................... 16

6.3 Dublin ................................................................. 16
6.3.1 James Hammerton, Postdoc ........................................... 16
6.3.2 Ronan Reilly (Coordinator) ........................................... 19
6.3.3 Related Dublin Researchers .......................................... 20
6.3.4 Training Activities ..................................................... 21
6.3.5 Network Collaboration ................................................ 21

6.4 ISSCO, Geneva ....................................................... 21
6.4.1 Adelina Hild, Postdoc ................................................ 21
6.4.2 Alexander Clark, Ph. D. Student ................................... 21
6.4.3 Susan Armstrong, Site Coordinator ................................ 22
6.4.4 Related Geneva Researchers ........................................ 22
6.4.5 Training activities ..................................................... 23

6.5 XRCE Grenoble ....................................................... 23
6.5.1 Nicola Cancedda, postdoc ............................................. 23
6.5.2 Christer Samuelsson and Eric Gaussier, coordinators .......... 24
6.5.3 Related research at XRCE ............................................ 25
6.5.4 Training activities ..................................................... 25

6.6 Groningen ............................................................ 25
6.6.1 Miles Osborne, Postdoc ............................................... 25
6.6.2 Stasinos Konstantopoulos, PhD student ........................................... 27
6.6.3 Susanne Schoof, PhD student ....................................................... 27
6.6.4 John Nerbonne, Coordinator ......................................................... 27
6.6.5 Related Groningen researchers .................................................... 28
6.6.6 Training activities ............................................................................. 28

6.7 Tübingen, Germany .............................................................................. 29

6.7.1 Franck Thollard, Postdoc ................................................................. 29
6.7.2 Alexander Clark, PhD. Student ......................................................... 30
6.7.3 Hervé Déjean, Postdoc .................................................................... 31
6.7.4 Dale Gerdemann, coordinator ........................................................ 31
6.7.5 Related Tübingen Researchers ........................................................ 32
6.7.6 Training activities ............................................................................. 32

A Shared Task Definition, Text Chunking .................................................. 35
1 Introduction

There are researchers working at all sites, mostly postdocs, and in some cases, Ph.D. students. Here we summarise the learning technologies used by each site:

<table>
<thead>
<tr>
<th>Site</th>
<th>Technology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antwerp</td>
<td>Instance-based Learning</td>
</tr>
<tr>
<td>Dublin</td>
<td>Neural Networks</td>
</tr>
<tr>
<td>Geneva</td>
<td>Stochastic Automata</td>
</tr>
<tr>
<td>Groningen</td>
<td>MDL, Random Fields (Maximum Entropy), ILP, Neural Networks</td>
</tr>
<tr>
<td>SRI</td>
<td>Genetic Algorithms, Maximum Entropy, ILP</td>
</tr>
<tr>
<td>Tübingen</td>
<td>Theory Refinement</td>
</tr>
<tr>
<td>Xerox</td>
<td>EBL</td>
</tr>
</tbody>
</table>

Erik Tjong Kim Sang and Stasinos Konstantopoulos maintain a very extensive project website at http://lcg-www.uia.ac.be/ containing notes on sites and on meetings, reports, data and programs.

2 Training

The LCG project continued to provide world-class training for its postdocs and graduate students.

- The Xerox Grenoble laboratory organised a four-day course in finite state methods given by Dr. Ken Beesley and Dr. Lauri Karttunen, both of the Xerox Research Laboratory in Grenoble. This was open to all TMR postdocs and predocs, and eight took part.

- At the joint CoNLL/LLL/ICGI meeting in Lisbon, the Portuguese Association for Artificial Intelligence organised a tutorial series on “Information Extraction and Learning from Text”. The eight talks concerned automatic thesaurus construction, part-of-speech tagging, extraction of patterns of words, alignment of parallel text and extraction of translation equivalents, document categorisation and other natural language processing tasks. Various techniques such as neural networks, inductive logic programming, statistics, memory-based language processing and symbolic adaptive parsing were explained. Five LCG members took part.

- Each of the sites continued to offer training to LCG postdocs locally as well as through tutorials at the meetings of professional societies (see Sections 6.5.1, 6.6.5) and to involve them crucially in providing training to beginning professionals. Details are in the section on the sites (Section 6).
3 Industrial Involvement

The industrial partner, Xerox, sponsored a visit to the Grenoble laboratories and conducted a tutorial in the use of some of their software. See Section 2. Project members Nicola Cancossedda and Christer Samuelson, both of Xerox, presented the LCG work at the Applied Natural Language conference in Seattle (Cancedda & Samuelsson 2000a).

Walter Daelemans (Antwerp) continued collaboration with S.A.I.L. trust, and also started a new project with E-Corporation (Alcatel) (see Section 6.1.2). Daelemans continued his active role in CLIF, *Computational Linguistics in Flanders*, a national research organisation for language processing. CLIF has university and industrial members.

Rob Koeling (SRI Cambridge) left LCG after July 2000 but continued at SRI for the rest of the year carrying out contract research. Koeling has accepted a permanent industrial position at NetDecisions, where he will apply his maximum entropy modeling.

Susan Armstrong (Geneva) continued work with Xerox and WIPO (World Intellectual Property Organisation) on the Automatic Terminology Extraction Project.

Hervé Déjean (Tübingen) has accepted a permanent industrial position at LCG partner Xerox’s Grenoble lab.

John Nerbonne and Gosse Bouma (Groningen) have begun an application in Text Classification in collaboration with BSC, the second largest Dutch company specialised in customer contact services.

4 Collaborations


Gertjan van Noord (Groningen) and Dale Gerdemann (Tübingen) continued their collaboration. This has resulted in two papers (reported in 2000) and two recent presentations which will be written up for publications. There are details in Section 6.7.4.

Erik Tjong Kim Sang and Walter Daelemans (Antwerp) hosted James Hammerton (Dublin) for a series of experiments aimed at understanding the degree to which neural network learning is hampered by sheer learning time limitations in dealing with large data sets. Tjong Kim Sang has done further joint work with two other project postdocs/predocs: Rob Koeling (Cambridge) on NP chunking, and Hervé Déjean (Tübingen) on NP chunking and while developing the CoNLL-2001 shared task.

Ronan Reilly (Dublin) served as outside examiner at Ivilin Stoianov’s dissertation defence
(Groningen) in March, 2000. Stoianov was earlier able to visit Reilly’s laboratory in the context of the LCG collaboration.

Nathan Vaillette (Tübingen) has agreed to visit Groningen to present his logical analysis of several extensions to FSA Toolkit, developed by Gertjan van Noord (Groningen).

Rob Koeling (SRI Cambridge) gave a tutorial on Maximum Entropy modeling in Tübingen.

5 Meetings

There were formal meetings in May and February and a large, informal gathering at the CoNLL/LLL/ICGI in Lisbon in Sept.

5.1 Grenoble, May 2000

The meeting was preceded by a four-day course in finite state methods given by Dr. Ken Beesley and Dr. Lauri Karttunen of the Xerox Research Laboratory in Grenoble. This was open to all TMR postdocs and predocs, and eight took part. The meeting continued the LCG tradition of an inviting an outstanding external speaker to present his work and to react to LCG work. Christo Samuelson, who organised the meeting, invited Colin de la Higuera, who is best known for his work in automata induction.

TMR-LCG Meeting Grenoble May 2000
Location: Xerox Research Centre Europe XRCE, 6 chemin de Maupertuis, 38240 Meylan (Grenoble), France.

Programme, Friday 19 May

<table>
<thead>
<tr>
<th>Time</th>
<th>Speaker/Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:30-10:15</td>
<td>John Nerbonne (Coordinator)</td>
</tr>
<tr>
<td>10:15-10:30</td>
<td>James Hammerton</td>
</tr>
<tr>
<td>10:30-11:00</td>
<td>Coffee</td>
</tr>
<tr>
<td>11:00-11:20</td>
<td>Anja Belz</td>
</tr>
<tr>
<td>11:20-11:40</td>
<td>Erik Tjong Kim Sang</td>
</tr>
<tr>
<td>11:40-12:00</td>
<td>Nicola Cancedda</td>
</tr>
<tr>
<td>12:00-14:00</td>
<td>Lunch</td>
</tr>
<tr>
<td>14:00-15:30</td>
<td>Colin de la Higuera (Invited speaker)</td>
</tr>
<tr>
<td>15:30-16:00</td>
<td>Coffee</td>
</tr>
<tr>
<td>16:00-16:10</td>
<td>Stasinos Konstantopoulos</td>
</tr>
<tr>
<td>16:10-16:20</td>
<td>Hervé Déjean</td>
</tr>
</tbody>
</table>
5.2 Tübingen, February 2001

The LCG project met for its prefinal meeting in Tübingen on Tuesday and Wednesday February 20-21, 2001.

Programme, Tuesday February 20

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>13:30-14:30</td>
<td>John Nerbonne (Coordinator)</td>
</tr>
<tr>
<td></td>
<td>Opening, Agenda, Goals</td>
</tr>
<tr>
<td>14:30-14:45</td>
<td>Walter Daelemans &amp; Erik Tjong Kim Sang</td>
</tr>
<tr>
<td></td>
<td>CoNLL CFP</td>
</tr>
<tr>
<td>14:45-15:00</td>
<td>Lisbon — CoNLL, ICGI, LLL</td>
</tr>
<tr>
<td>15:00-15:15</td>
<td>Annual Report</td>
</tr>
<tr>
<td>15:15-15:45</td>
<td>Final Meeting — Toulouse</td>
</tr>
<tr>
<td>15:45-17:15</td>
<td>Postdoc/Predoc Reports</td>
</tr>
<tr>
<td></td>
<td>Belz, Hammerton, Konstantopoulos, Schoof</td>
</tr>
</tbody>
</table>

Wednesday February 21

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:00-9:15</td>
<td>Coffee</td>
</tr>
<tr>
<td>9:15-10:45</td>
<td>postdoc reports</td>
</tr>
<tr>
<td></td>
<td>Thollard, Clark</td>
</tr>
<tr>
<td>10:45-11:15</td>
<td>Coffee</td>
</tr>
<tr>
<td>11:15-12:45</td>
<td>postdoc reports</td>
</tr>
<tr>
<td></td>
<td>Tjong Kim Sang, Dejean, Cancetta</td>
</tr>
<tr>
<td>12:45-14:15</td>
<td>Lunch</td>
</tr>
<tr>
<td>14:15-15:45</td>
<td>Rens Bod, Amsterdam</td>
</tr>
<tr>
<td></td>
<td>Invited Lecture, Discussion</td>
</tr>
<tr>
<td>15:45-16:15</td>
<td>Coffee</td>
</tr>
<tr>
<td>16:15-17:15</td>
<td>Other issues, Quo vadis, LCG? (follow up)</td>
</tr>
<tr>
<td>17:15</td>
<td>Closing</td>
</tr>
</tbody>
</table>

5.3 Lisbon, September 2000

LCG participants were actively involved in the following triple of conferences, co-located in Lisbon, Portugal, September 11th–14th, 2000. On Sept. 14 the LCG project organized a
shared-task session on phrase chunking.

- a) 5th International Colloquium on Grammatical Inference (ICGI),
- b) 2nd Learning Language and Logic Workshop (LLL) and
- c) 4th Conference on Computational Natural Language Learning (CoNLL).

This triple conference event attracted about 100 participants from 22 countries from all over the world. 12 participants came from each of the countries Japan, France and Spain, which made them the leading participating nations. Most accepted papers were presented orally.

The program was as follows:

- Sept. 11th: 11 ICGI talks scheduled from 10:00 till 17:30.
- Sept. 12th: 9 ICGI talks scheduled from 9:00 till 15:30; one talk was skipped due to the absence of the speaker(s).
  In the late afternoon, a nice city tour had been organised.
- Sept. 13th: 4 ICGI talks; 6 CoNLL talks (including an invited lecture by Dan Roth) and 1 LLL talk scheduled from 9:00 till 15:30. Then, a bullet presentation (12 talks) and the corresponding CoNLL poster session followed.
  In the evening, a beautiful banquet had been organised at the museum of electricity.
- Sept. 14th: 3 shared LLL/CoNLL talks from 9:30 till 12:00. Sessions were then separated: 5 talks for the LLL conference and 4 CoNLL talks followed by a shared-task session organised by the TMR LCG project. The shared task concerned machine learning of chunking, i.e., syntactic constituent detection and labeling in free text).

All this yielded a stimulating atmosphere with many interdisciplinary impacts. It was felt that a closer collaboration of people working in pattern recognition, artificial intelligence, natural language processing and formal language theory could produce many new insights, ideas and applications.

Immediately after the conferences, a tutorial series on “Information Extraction and Learning from Text” was sponsored by the Board of the Portuguese Association for Artificial Intelligence. The eight talks concerned automatic thesaurus construction, part-of-speech tagging, extraction of patterns of words, alignment of parallel text and extraction of translation equivalents, document categorisation and other natural language processing tasks. This way, various techniques such as neural networks, inductive logic programming, statistics, memory-based language processing and symbolic adaptive parsing were explained.

For further references, consult:
5.4 ICGI-2000, Lisbon, September 2000

ICGI-2000 in Lisbon was the fifth in a series of successful biannual international conferences on the area of grammatical inference. Grammatical inference has been extensively addressed by researchers in information theory, automata theory, language acquisition, computational linguistics, machine learning, pattern recognition, computational learning theory and neural networks. This colloquium aimed at bringing together researchers in these fields. Previous editions of these meetings were held in Essex, U.K.; Alicante, Spain; Montpellier, France; and Ames, Iowa, USA.

5.5 CoNLL, Lisbon, September 2000

The CoNLL was the fourth in a series of meetings organised by SIGNLL, the ACL's Special Interest Group on Natural Language Learning. Previous meetings took place in Madrid, Sydney and Bergen. The focus of the CoNLL workshop is to address issues in machine learning of language, including issues that are not regularly discussed at computational linguistics meetings, such as computational models of human language acquisition, models of the evolution of language etc. The CoNLL had contributions from America (7), Europe (27) and Asia (3). 25% of the program stemmed from LCG work!

A particularly gratifying development was the participation of many (five) non-LCG groups in the LCG shared task. This demonstrates that LCG is impacting technical work beyond the range of the seven partners.

The appendix to this report contains the definition of the shared task and a list of papers presented in connection with it.

5.6 LLL, Lisbon, September 2000

The purpose of LLL was to provide a forum for discussion on all aspects of learning language in logic. It was the follow-up of the first LLL workshop held in Bled (Slovenia) in 1999; at that time, it was co-located with ICML-99 (International Conference on Machine Learning). Relational learning and logic-based learning have proved their capacity to learn complex structured knowledge from structured data and explicit background knowledge. Compared to data analysis, some of the major advantages are a better means to express the representation; a method that is easier to understand; and a comprehensible learning result.
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, 2000 – March 31, 2001.

6.1 Antwerp

This site employs one postdoc. Apart from his research progress report, this section also contains an overview of the work of the local coordinator and some notes about the training activities at this site.

6.1.1 Erik Tjong Kim Sang, Postdoc

In the third project year (April 1, 2000 - March 31, 2001), Erik has mainly worked on general chunking and parsing. He also finished his earlier work on system combination approaches to noun phrase chunking. This work has been reported on at papers presented at NAACL 2000 and Coling 2000.

The general chunking work consisted of the shared task for CoNLL-2000. Here, Erik applied an extension of his NAACL 2000 method for NP chunking to general chunking. His system finished third in a field of 11 participants. Erik also coordinated this shared task. Among the eleven participants were another three from our project: Hervé Déjean (Tübingen), Rob Koeling (Cambridge) and Miles Osborne (Groningen).

Erik built a basic memory-based parser by combining his work on NP bracketing (CoNLL-99) and arbitrary chunking (CoNLL-2000). Although the two base applications perform well, the performance of the parser is disappointing (F=80 compared with F=90 for the best statistical parsers). This work was reported on at CLIN 2000 and a paper describing the work has been submitted for the proceedings of this conference.

During this year, Erik has done joint work with three project postdocs/predocs: Rob Koeling (Cambridge) on NP chunking, James Hammerton (Dublin) in a comparative study of connectionist networks and memory-based learning, and Hervé Déjean (Tübingen) on NP chunking and while developing the CoNLL-2001 shared task. Furthermore, he wrote a collection paper on part of his PhD thesis work with John Nerbonne (Groningen). Other joint papers were written with researchers of the University of Tilburg (The Netherlands), Bar-Ilan University (Israel), the University of Illinois (USA) and local colleagues from the University of Antwerp. James Hammerton, TMR postdoc in Dublin, visited the Antwerp research group for six weeks in the period May-July 2000.
6.1.2 Walter Daelemans, Coordinator

CNTS (co-directed by Daelemans and Gillis) has continued investing a large research effort in Machine Learning of Natural Language. In addition to the LCG project, in 2000, research continued on the application of classifier combination techniques and evolutionary computing to language and speech technology problems (funded by S.A.L. trust; Anne Kool, Jakub Zavrel, and Walter Daelemans). More work on classifier combination, meta-learning, and rule induction in the context of grapheme-to-phoneme conversion was done in the Linguaduct project (Veronique Hoste, Walter Daelemans, Steven Gillis). Additional relevant work includes that of Guy De Pauw (evolution of grammars), Masja Kempen (machine learning of models of language acquisition), and the project on computational psycholinguistics (computational models of human language learning and processing; Gert Durieux, Helena Taelman, Evelyn Martens, Steven Gillis, Walter Daelemans). Topics studied here include syllable structure acquisition, word stress acquisition and German plural acquisition and processing. They also applied machine learning techniques for boosting annotation speed and accuracy in the Dutch-Flemish Corpus Spoken Dutch project (for tagging and phonetic transcription; Hoste, Depoorter, Zavrel, Daelemans, Gillis).

Two new related projects started in 2000:

- Funded by E-Corporation (an Alcatel daughter), Eric Van Horenbeeck started a project on the application of machine learning techniques to the problem of topic detection in newspaper text.

- A large Flemish project in cooperation with computational linguistics and speech technology groups in Leuven and Gent was started on the Automatic Transcription and Normalisation of Speech. Target applications in this project for CNTS are phoneme-to-grapheme conversion of out-of-vocabulary items, and automatic subtitling.

CNTS coordinates CLIF (Computational Linguistics in Flanders, a National Science Foundation sponsored research association), and members of CNTS are active in various advisory or management functions of national, transnational (Dutch-Flemish), and European government-initiated initiatives on language and speech technology.

Apart from the people mentioned, several researchers from Tilburg University (ILK group, co-directed by Van den Bosch and Daelemans) were still active in LCG and related projects, especially Sabine Buchholz and Jorn Veenstra.

Relevant publications of Daelemans and other CNTS participants about machine learning of natural language can be found in the bibliography.

6.1.3 Training Activities

Erik has participated in the tutorial in finite state technology methods that was offered to the TMR postdocs in May 2000 by Ken Beesley and Lauri Karttunen of Xerox Greno-
ble. He gave three undergraduate lectures in Antwerp and two lectures in Ypre (S.A.I.L). The lectures covered statistical and connectionist natural language processing. Erik also supervised a student who wrote a Masters thesis on connectionist language processing.

6.2 SRI International, Cambridge, UK

This site employs 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.

6.2.1 Summary of LCG Project 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 addition, the seminar series that started in March 2000 continued through Spring and Summer 2000. Project researchers also maintain extensive project-related web pages, and attended project-related conferences, workshops, seminars and meetings.

The LCG group at SRI organise 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 and summer 2000 a special seminar series on Statistical Methods for NLP and NLL took place. This consisted 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 international conferences and meetings, including the LCG project meeting in Grenoble (May 2000), Gotalog (May 2000), Coling-2000 and SIGPhon-2000 (July), and the LCG project meeting in Tübingen (February 2001). They also presented talks on results achieved for LCG learning tasks (for details, see below).

Collaboration between the local full-time LCG researchers and associated project members, in particular at the University of Cambridge Computer Laboratory, continues to form part of the Cambridge LCG activities. Members of the Computer Laboratory have been attending Local LCG Seminars and presenting talks.

Plans for the next year focus on the completion of ongoing research, and the publication of papers reporting the results.
6.2.2 Anja Belz, Postdoctoral Researcher

During the reporting period, Anja Belz continued to work with treebank grammars and probabilistic context-free grammars (PCFGs) with local structural context (LSC). This project 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.

The first stage focused on investigating different ways of creating tree bank grammars, testing these on different parsing tasks and comparing the results to existing research. The second stage looked at using tree bank grammars as a starting point for automatic grammar construction. Different types of LSC (grammatical function, parent node information, and depth of embedding) were incorporated in tree bank PCFGs and their effect on parsing performance was tested. Results show that all types of LSC investigated improve parsing performance, with parent node information the most immediately useful. The third project stage concentrates on automatic search methods for optimising LSC-PCFGs for a given task in terms of grammar size and parsing performance (F-Score). Results for the first two stages, and preliminary results for the third were reported in a paper at Corpus Linguistics 2001 (Belz, 2001).

Anja Belz 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).

Apart from conducting research on project-related subjects, Anja Belz maintained and extended the local project web pages, and organised a series of seminars on subjects related to TMR LCG research which included reading group sessions, presentations by local researchers (including tutorials) and invited talks by researchers from other organisations.

Plans for the next year focus on completing stage 3 of the LSC-PCFG project, writing a detailed technical report on the entire project, and submitting a shorter version of the technical report to a major international journal.

6.2.3 Rob Koeling, Doctoral Researcher

Rob Koeling continued to work on applying Maximum Entropy models to the shared tasks described previously. Rob Koeling attended the LCG project meeting in Grenoble (May
2000) and the Gatalog conference in Gothenborg (May 2000). A talk was presented at the project meeting in Grenoble and a poster was presented at Gatalog. The poster presentation at Gatalog 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 (among others) several LCG project members. Koeling (2000b) was his contribution to the shared task defined for the CoNLL 2001 workshop of finding arbitrary chunks in data.

As part of the spring and summer series of seminars at SRI, Rob Koeling gave a Maximum Entropy tutorial for local TMR-LCG researchers and some interested members of the Cambridge NLP community.

Koeling's involvement with TMR-LCG ended 1 August 2000.

6.2.4 David Milward, Coordinator

David Milward has continued to supervise Anja Belz and, until August 2000, Rob Koeling, and to attend local reading groups on machine learning. 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 Involvement of other Researchers

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.

Several members of the Cambridge NLP community regularly attended LCG seminars, including Dr Ted Briscoe, Dr Ian Lewin, Dr Richard Tucker, Sylvia Knight, Aline Villavicencio, Ben Waldron, and other doctoral researchers at the Computer Laboratory.

Invited speakers last year included Dr Mark Hepple (Sheffield) who gave a presentation on treebank grammar research in Sheffield, and Dr Miles Osborne (former LCG postdoc in Groningen, now lecturer in Edinburgh) who presented his paper *Estimation of Stochastic Attribute-Value Grammars using an Informative Sample*, and Aline Villavicencio who presented her paper *The Acquisition of a Unification-Based Generalised Categorial Grammar*.

6.2.6 New Publications

1. (Belz 2001)
2. (Belz 2000)
3. (Koeling 2000a)

4. (Tjong Kim Sang, Daelemans, Déjean, Koeling, Krymolowski, Punyakanok & Roth 2000)

5. (Koeling 2000b)

6.2.7 Training Activities

Anja Belz attended the LCG project meetings in Grenoble (May 2000), including the Finite-State Methods course. She also organised a series of seminars on subjects related to TMR LCG research which included reading group sessions, presentations by local researchers (including tutorials) and invited talks by researchers from other organisations.

Rob Koeling also attended the LCG project meetings 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 local TMR-LCG researchers and some interested members of the Cambridge NLP community. He also gave this tutorial to the Tübingen LCG group.

6.3 Dublin

This site employs one postdoc. This section contains an overview of his research progress and training activities since April, 2000 as well as the activities of the local coordinator and other people working at this site.

6.3.1 James Hammerton, Postdoc

At the time of the previous report, James was investigating the use of the SARDSRN network for the noun-phrase bracketing task. Since then changes to the way the problem was presented to the network, plus addition of connections directly from the inputs to the outputs improved on the results reported previously. As before the network outputs a representation of the bracketed sequence with place markers for the word. However, now the network outputs partial results. Whenever a noun-phrase closes it outputs the bracketed sequence thus far. Otherwise no targets are presented, thus the network concentrates on learning only the key outputs. The best result obtained was that a SARDSRN with 2, 49-unit SARDNETs, 144 hidden and context units, 32 inputs and 40 outputs was trained on all the sentences of length < 10 in the training corpus. It was tested on all 184 sentences of length < 10 in the testing corpus. It learned to process 550 of the training sentences perfectly, yielding a training set F-score of 96.2 and a testing set F-score of 49.77. It only processed 29 of the test sentences perfectly. Training took 26 hours and only 3 out of 5 runs approached this level of performance.
While this was a significant improvement over the previous performance, it was felt that the training was too intensive and thus an alternative approach was needed. It was decided to investigate whether a new neural network architecture known as Long Short-Term Memory (LSTM) (Hochreiter & Schmidhuber 1997) would perform better than SARDSRN. LSTM is explicitly designed to tackle problems of recurrent networks forgetting information presented to them from several time-steps ago.

The main innovation in LSTM is a new style of hidden layer. LSTM networks have only 3 layers, with the hidden layer being a recurrent layer. The hidden layer consists of a set of memory blocks each containing one or more memory cells and an input and output gate. A single-celled memory block is depicted in Figure 1. The cell is a linear unit with a self connection whose weight holds the value 1. If there is no other input to the cell, then it simply holds its current value indefinitely over time. The inputs to the cell are multiplied by the activation of the input gate. Thus the input gate controls what information gets into the cell. Similarly the outputs from the cell are multiplied by the activation of the output gate. The output gate thus controls when the cell's information flows out to the outputs. The input and output gates thus learn to hold onto cell states for arbitrary periods of time. LSTM is trained via an algorithm combining back propagation through time with real-time recurrent learning. LSTM has been demonstrated to be capable of remembering information over a period of 1000 time-steps on artificial tasks. It was hoped that it would provide improved performance over the SARDSRN.

Currently, the best result for LSTM is that a network with 32 inputs, 15 blocks of 10 memory cells, 40 outputs, employing the same input and output representations as SARDSRN and with the inputs connected directly to the outputs learned to process 736 of the training sentences of length < 10 perfectly, and achieved an F-score of 100 on the training set. On the sentences of length < 10 in the testing set it processed 46 of them perfectly achieving an F-score of 62.79. The training took 19 hours, and was more reliable – most runs were
<table>
<thead>
<tr>
<th></th>
<th>Train correct, F-score</th>
<th>Test correct, F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>MBL 0-0</td>
<td>76, 82.19</td>
<td>7, 69.57</td>
</tr>
<tr>
<td>MBL 1-0</td>
<td>181, 57.76</td>
<td>3, 17.07</td>
</tr>
<tr>
<td>MBL 2-0</td>
<td>96, 25.88</td>
<td>2, 15.05</td>
</tr>
<tr>
<td>LSTM</td>
<td>736, 100</td>
<td>46, 62.79</td>
</tr>
<tr>
<td>SARDSRN</td>
<td>551, 96.89</td>
<td>29, 49.77</td>
</tr>
<tr>
<td>ErikMBL 1-0</td>
<td>n/a, 97.89</td>
<td>n/a, 65.92</td>
</tr>
</tbody>
</table>

Table 2: Comparison with MBL. The numbers given are the number of sentences correct and the F-score.

giving results close to these.

While this is an improvement, it is still felt that the training is too intensive – using the same network and training regime to learn to process all the 3456 training sentences of length < 20 would take over 7 days by extrapolation, and it is likely a larger network would be needed. For this reason a rethink of the approach is being investigated.

The poor generalisation seen with both LSTM and SARDSRN may be due to the small sample of the data set being used. In order to test this hypothesis it was decided to compare the performance of LSTM and SARDSRN with Memory-Based Learning. An MBL system was created which used the same inputs and produce the same outputs as the neural networks and was trained with progressively increasing amounts of left context. The F-scores and number of sentences correct were then computed for both the training and testing sets. The results are given in Table 2.

As can be seen, both MBL and the networks generalise poorly, in terms of the number of sentences processed perfectly. However the networks do far better than MBL on this measure.

Curiously, MBL with only the current word in the input window outdoes all the rest on the test set F-score. “ErikMBL” is a run performed by Erik Tjong Kim Sang’s MBL system which recursively finds noun phrases started with the base noun phrases. LSTM is the best performing network and is at least in the same ball park as MBL. For this reason, SARDSRN is being abandoned and ways of improving LSTM’s performance explored.

Currently, the training places some unnecessary burdens on the network. Firstly, by being required to produce a representation of the entire sequence at the output layer, the hidden layer pattern in successful training is required to be translatable to any symbol in the sequence in a single step. Secondly, the role of a set of output units representing a symbol in the bracketed sequence varies from sequence to sequence. In some sequences it will represent a word, in others it will represent an opening or closing bracket. The output representation was chosen to facilitate possible holistic processing, however this can be achieved via developing a hidden layer representation instead, which may be easier for the network.
For the reasons given above a new output representation is being investigated. The network will be presented with each sentence, one word and POS tag at a time, in 2 passes. On the first pass no target outputs are produced, then on the second pass for each word, the network outputs the number of opening and closing brackets at that word. With this new approach the hidden layer is no longer constrained as before, and each output pattern plays the same role across all sentences. It is hoped that this will achieve easier, quicker training and better performance. This can be modified so that after the 1st pass, the network is trained to output the same information but without the words being presented. Trained this way, a holistic representation of the bracketed sequence would be developed at the hidden layer without the burdens placed by the earlier representation.

Another modification addresses the possibility that the small sample size from the data may itself be a problem. Instead of restricting the length of the sentences, the training sentences will be sampled from section 15 to 18 of the journal with no restriction on length. This may improve the learning by improving the quality of the data being used.

- Hammerton attended the “Computational Natural Language Learning Workshop (CoNLL-2000)” held in Lisbon, Portugal.
- He presented, as co-author, a paper entitled “Holistic Unification and the Bi-coding RAAM”, P. Heffernan & J.A. Hammerton at the Intelligent Systems and Architectures conference (ISA 2000) held at the University of Wollongong, Wollongong, Australia, 12-15 December 2000. This paper was based on the Paul Heffernan’s final year project which Hammerton proposed and co-supervised with Ronan Reilly as reported in the previous report. Paul was unfortunately unable to make it to the conference himself, due to work commitments.

At the last project meeting in Tübingen Hammerton volunteered to edit a special issue of a journal on machine learning approaches to shallow parsing. It is intended that project members submit their work to this journal, although it will also be open to non-project members.

### 6.3.2 Ronan Reilly (Coordinator)

Ronan directs the MA/MSc in Cognitive Science which has involved contributions from various affiliates of the LCG project at UCD. This has had at least two important consequences; the first and most important one is that it has provided the students involved with a strong grounding in interdisciplinary research with a significant focus on language. The second consequence is that it has made those involved in teaching the course aware of each other's work, and thus laid the groundwork for future research collaboration.

Ronan gave several talks in this period:


Ronan was also a PhD examiner for a project partner at the University of Groningen (Ivilin Stoianov) and had several publications, (Akhtar & Reilly 2001 (forthcoming), Kechadi & Reilly 2000, Mackey & Reilly 2001 (forthcoming), Neouva & Reilly 2001 (forthcoming), Reilly 2000, Reilly & Mackey 2001 (forthcoming)).

### 6.3.3 Related Dublin Researchers

Members of staff involved in the Cognitive Science MSc/MA included Fred Cummins, Julie Berndsen, Arthur Cater, Gregory O'Hare and Mark Keane.

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 the way the output from the network was being represented.

Talks by department members included the following:

- Fred Cummins. Invited speaker at Department of Linguistics and Finnish Institute, Stockholm University, Spring 2000.

6.3.4 Training Activities

- James Hammerton attended the course Xerox ran on their Finite State Toolkit for NLP in Grenoble in May 2000.
- He proposed and co-supervised (with Nick Kushmerick) a summer project where the student, Paul Heffernan, investigated whether an inductive logic learner called FOIL could learn to perform unification on the same data as used in the holistic unification project. It was found that it could, although the training was very intensive.

6.3.5 Network Collaboration

A research visit was arranged for the 28th May to 9th July 2000 at the University of Antwerp. During this period Hammerton collaborated with Erik Tjong Kim Sang on producing a hybrid neural network/MBL system. This system employed a self-organising map (SOM) to select a subsection of training items for comparison with a test item to investigate whether this would allow faster processing through a reduced number of comparisons without reducing performance. This system was implemented and experiments were run which showed that it could achieve a level of performance close to that of MBL but using a far smaller number of comparisons. A paper on this is currently being prepared and it is hoped it will be accepted for the CoNLL 2001 workshop in Toulouse.

6.4 ISSCO, Geneva

This site employed one postdoc and more recently a PhD student. 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.

6.4.1 Adelina Hild, Postdoc

During the project period, Adelina Hild continued her work on the linguistic analysis of the acquisition and processing of NP’s with particular emphasis on recursive NP’s. 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 Alexander Clark, Ph. D. Student

Alexander Clark joined ISSCO on October 1st, 2000. He is currently in the third year of his PhD. at the University of Sussex, under the supervision of Dr Bill Keller. The provisional title of his thesis is "Unsupervised Language Acquisition: Theory and Practice". 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. Shortly before joining the project he attended CoNLL-2000 where he presented a short paper entitled "Inducing Syntactic categories using context distribution clustering".

He has been working on the supervised and unsupervised learning of morphology using Pair Hidden Markov Models, a model used for aligning DNA sequences in computational biology. These models can be used to learn any sort of finite-state string transductions but seem especially well-suited to learning morpho-phonology. He gave an invited talk on this subject to the Natural Language and Computational Linguistics seminar group at the University of Sussex, on December 1st 2000, and has submitted a paper to the student session at EACL 2001.

Alexander's other research addresses the issue of unsupervised learning of syntax using the technique of distributional clustering to induce a stochastic context-free grammar. He gave a talk about this at the Tübingen meeting. Unsupervised methods have several advantages – there is a much larger amount of data available, and one is not tied into the particular set of (potentially sub-optimal) analytical decisions chosen by the annotators of the corpus.

### 6.4.3 Susan Armstrong, Site Coordinator

In addition to supervising the work carried out in the TMR/LCG project by the collaborators at ISSCO, she has conducted research and given presentations in automated terminology acquisition. In collaboration with Xerox, Grenoble and WIPO (World Intellectual Property Organisation) she continues work on the Automatic Terminology Extraction Project. Another project in progress, related to the identification of terminology, is the automatic clustering of similar texts.

### 6.4.4 Related Geneva Researchers

Colleagues at ISSCO participated in the EU funded 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.
6.4.5 Training activities

Two interns collaborated in 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.

6.5 XRCE Grenoble

This site employs one postdoc. This report contains an overview of his research and training activities and the work of the coordinator and others at this site.

6.5.1 Nicola Cancedda, postdoc

TMR-LCG research at Xerox was focused on two distinct problems:

- Extending techniques previously developed for grammar specialisation to the problem of grammar adaptation;
- Annotating a bilingual parallel corpus using a monolingual parser and a bilingual dictionary.

A short description of the work conducted on each of the two problems follows.

Grammar specialisation is intended as the problem of trading reduction in grammar coverage for reduction in grammar ambiguity in the most favourable way. Cancedda’s work has led to a formulation of the problem as a local optimisation problem in the space of grammars, in which search is made feasible by a new representation for tree banks called \textit{folded tree banks}. Distinct aspects of this work were presented at the conference on Applied Natural Language Processing (ANLP 2000), in Seattle (Cancedda & Samuelsson 2000b), and at the workshop on Computational Natural Language Learning (CoNLL-2000), in Lisbon (Cancedda & Samuelsson 2000a).

In the period covered by this report Cancedda investigated the theoretical aspects of an extension to the more general problem of \textit{Grammar adaptation}. Grammar adaptation differs from grammar specialisation in that alterations to the initial grammar are considered which are no longer limited to reductions in coverage, but could instead lead to an \textit{increase} in coverage. In particular, he considered the combination \textit{Inductive Chart Parsing} (Cussens & Pulman 2000a, Cussens & Pulman 2000b) with a representation derived from folded tree banks, the \textit{folded chart banks}. Whereas in folded tree banks complete parse trees are folded
on a special purpose representation of the grammar, in folded chart banks it is the charts produced by a bottom-up chart parser to be folded onto the same grammar representation.

Folded chart banks hold the premises to allow a much more efficient assessment of the effects of modifying a grammar than the more conventional representation adopted so far in inductive chart parsing. Planned experiments will show whether folded chart banks can allow scaling up from the toy examples studied so far in inductive chart parsing to large-scale grammars and corpora.

Most of the work conducted within the LCG project concerns learning stochastic or symbolic models for some form of syntactic processing. An important obstacle to the proposed approaches and to statistical parsers in general lies in the necessity of an extensive annotated corpus for training. This obstacle is even more serious whenever languages other than English are considered. A possible solution to this problem consists of using an existing parser developed for a language to build the training examples from which to induce a parser for a different language. More specifically, the scenario that he has been investigating is the following.

Let $C_1$ be an available chunker for language $L_1$, and let $BC_{12}$ be a bilingual $L_1/L_2$ corpus, aligned at the sentence level. Moreover, let $BD_{12}$ be a bilingual $L_1/L_2$ dictionary. The following steps are performed:

1. chunk $BC_1$, the part of $BC_{12}$ in language $L_1$, using $C_1$;
2. align $BC_{12}$ at the word level relying on the bilingual dictionary $BD_{12}$;
3. project the chunks obtained at step 1 through the alignment obtained at step 2.

If the chunking can be projected with sufficient accuracy, than many available techniques can be used to train a new chunker for language $L_2$.

Preliminary experiments have been conducted using the chunking component of the Xerox Incremental Parser (XIP) for French, and the bilingual pair French/Italian of the ELRA MLCC 1.0 parallel corpus, aligned at the sentence level using existing tools.

### 6.5.2 Christer Samuelsson and Eric Gaussier, coordinators

Christer Samuelsson acted as site coordinator for the project until July 2000. In the period covered by this report, he conducted theoretical research on stochastic dependency parsing as well as a more applied effort aiming at improving optical character recognition using stochastic modeling.

Eric Gaussier replaced Christer Samuelsson as site coordinator in July 2000. Apart from supervising the work of Nicola Canciedda, he worked on unsupervised methods for prepositional phrase attachment resolution, with a focus on the combination of more reliable
and less reliable sources of information. Reliable information was extracted by means of regular expressions applied to the output of a chunker. Applying these patterns allowed the disambiguation of the concerned attachments with 0.95 precision. This information was then complemented by less reliable information through a windowing approach collecting all the pairs occurring within k elements. The influence of each source of information was tested within a probabilistic framework, relying on two models, a parametric model based on the multinomial distribution, and a distribution-free model the parameters of which are estimated through the EM algorithm. Preliminary results show the importance of reliable information in the PP attachment resolution task, but also suggest that combining the two sources in a motivated way can improve the system. Lastly, by comparing this approach to a baseline relying only on subcategorization frames encoded in our lexicon, a full justification to the use of machine learning techniques in PP attachment resolution was obtained.

6.5.3 Related research at XRCE

Many web sites automatically generate pages from underlying databases, often in response to user queries. The information contained in the database is presented according to some more or less rigid format. The structure of this format can be conveniently captured in a grammar. A wrapper is a program which, relying on such a grammar, can extract the tuples in the original database from the formatted page. Boris Chidlovskii developed a tool for inducing the formatting grammar from labeled examples in a semi-automatic way.

6.5.4 Training activities

Nicola Canceddà attended the ELSNET summer school on Text and Speech Triggered Information Access (TeSTIA '2000) in Chios, Greece. Moreover, he attended tutorials on “Language Modeling” and on “Machine Translation” at ANLP'2001. Nicola Canceddà also co-animals a reading group on Machine Learning.

6.6 Groningen

This site employs a 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 Miles Osborne, Postdoc

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 current position on the 1st of October 1998. He accepted a post as lecturer at the University of Edinburgh as of 1 Jan 2000, but continued on the LCG project for two months in the summer of 2000.

LCG related research activities in the third year were as follows:

- Induction of DCGs modelled as random fields. This is ongoing work, and follows 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 will perform 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).

  To date, a superset of DCG rules (16k) have been acquired, and existing iterative scaling code is being 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 published in (Osborne 2000) COLING 2000 (Saarbrücken, August 2000) in Osborne M., Estimation of Stochastic Attribute-Value Grammars using an Informative Sample. A journal length treatment has been submitted to Natural Language Engineering.

- As a by-product, random-field modelling should lead to competitive parse selection results. Results on parse selection have been presented at CoNLL-2000 (Lisbon, September 2000) along with Tony Mullen in T. Mullen and M. Osborne, Overfitting Avoidance for Stochastic Modelling of Attribute-Value Grammars. Tony Mullen also presented a paper at the student ACL session in Hong Kong, 2000.

- In CoNLL-2000, Miles also presented his Shallow Parsing as Part-of-Speech Tagging paper.

- In conjunction with Rob Malouf, Miles Osborne has begun investigating ensembles of Maximum Entropy / Minimum Divergence Models for Attribute-Value Grammars. This will result in a submission to ACL 2001.

- 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 Stasinos Konstantopoulos, PhD student

Stasinos is using the Aleph Inductive Logic Programming System developed in Oxford (see http://oldwww.comlab.ox.ac.uk/oucl/groups/machlearn/Aleph/aleph.html) to induce Dutch Phonotactics from the Dutch section of the CELEX corpus. He is currently experimenting on the effect of background knowledge in the quality of the rules constructed, and more specifically with the part of the background dealing with the segment features available to the learner. The background knowledge setups compared range from simple, language-independent phonetic descriptions of the Dutch segment inventory to more complex ones that are motivated by Dutch phonological processes. Results of this research have been accepted for the student session of the 13th European Summer School on Language, Logic and Information (ESSLLI 2001) that will take place in Helsinki in August 2001. (see http://www.helsinki.fi/esslli/ for more information on ESSLLI 2001.)

Non-LCG academic activities include maintaining the HP-UX port of the YAP Prolog System for which Aleph is written.

6.6.3 Susanne Schoof, PhD student

Susanne Schoof has joined the project in June, 2001 as a beginning PhD student. She will begin with grammar and corpus techniques.

6.6.4 John Nerbonne, Coordinator

John Nerbonne has continued investigations into the application of unsupervised learning to the problem of dialect classification with (a) Wilbert Heerenga, John Nerbonne and Peter Kleiweg Validating Dialect Comparison Methods submitted to Wolfgang Gaul and Gerd Ritter (eds.)


He has also done work on the subjects of phonological learning (Tjong Kim Sang and John Nerbonne, Learning the Logic of Simple Phonotactics, In: James Cussens and Saso Dzeroski (eds.), Learning Language in Logic. LNAI1925, 2000)

Together with Gosse Bouma he has begun a project on automatic email classification which makes extensive use of machine learning techniques.
6.6.5 Related Groningen researchers

Gertjan van Noord is the project manager of the *Algorithms for Linguistic Processing* project. ALP focuses on two crucial problem areas in computational linguistics: problems of processing efficiency and ambiguity. For the problem of efficiency grammar approximation techniques are investigated, whereas a number of grammar specialisation techniques (including machine learning methods) are tried for the ambiguity problem. See http://www.let.rug.nl/~vannoord/alp/ for more information on ALP.

Iyilin Stoianov successfully defended his PhD thesis on *Connectionist Lexical Processing* in March 2001. The parts of his work that are of most interest to LCG are chapters 4 and 5 of his thesis on the application of Neural Networks on Phonotactics and Grapheme-to-Phoneme conversion.

Tony Mullen is a Groningen PhD student who has been collaborating with Miles Osborne in a project aimed at locating where overfitting is most likely, and most damaging. Results on parse selection have been presented in T. Mullen, *Overfitting Reduction through Feature Measuring for Maximum Entropy-based Parse Selection*, ACL 2000 (Hong Kong) and in T. Mullen and M. Osborne *Overfitting Avoidance for Stochastic Modeling of Attribute-Value Grammars*, CoNLL-2000 (Lisbon, September 2000). He attended tutorials at the first of these meetings.

Rob Malouf joined the Groningen group in July, 1999 as part of a university project focused on computational modelling of behaviour, 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 work on efficient processing techniques (Malouf, Carroll & Copestake 2000) and Maximum Entropy modelling (*A toolkit for robust and efficient maximum entropy language modelling*, CLIN, November 2000). He was able to attend tutorials at the NAACL/ANLP meeting.

Gosse Bouma is a permanent staff member in Groningen who has attended LCG meetings and who has applied learning to the problem of grapheme-to-phoneme conversion (Bouma 2000). He and John Nerbonne collaborate with BSC (a customer contact company) on an application of LCG techniques to automatic email answering.

6.6.6 Training activities

Stasinos Konstantopoulos is assisting Gosse Bouma by giving tutorials for a course in Natural Language Processing. (March - June 2001, see http://www.let.rug.nl/~gosse/nlp1/ for more details.) He will also attend the 13th European Summer School in Logic, Language and Information (Helsinki, August 2001) where he has had a paper accepted for the student session. He has also 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. He has also attended various talks
on subjects of general interest (e.g. the weekly linguistics Colloquium in Groningen).

6.7 Tübingen, Germany

This site has employed one new full-time postdoc (Franck Thollard) and one new part-time predoc (Alexander Clark). The remainder of Clark's time is spent at the LCG consortium partner ISSCO in Geneva. Hervé Déjean will leave TMR-LCG 31 March 2001. The participations of Franck Thollard, Alexander Clark, Hervé Déjean and Dale Gerdemann are presented here, as well as related work of other researchers in Tübingen. The last section presents common training activities.

6.7.1 Franck Thollard, Postdoc

This section presents Franck Thollard’s participation in the project since he joined it on September 1st 2000. The first part outlines his background and is followed by a part describing his participation in the project including an outline of the conferences and seminars he attended as well as details on his collaborations and research activities.

Franck Thollard defended his PhD dissertation on the 10th of July, 2000. His work aimed at modeling natural language by way of probabilistic automata. The final goal was to use the automata in continuous speech recognition.

A new algorithm (namely MDI) was proposed in his PhD that produced better performance than the one produced by the classic algorithms. These automata were later on plugged into the speech recognition system of France Telecom research lab. See below for more details.

Another research axis was the theoretical aspect: it was shown in his PhD that probabilistic automata can be inferred in the identification in the limit paradigm.

Franck Thollard’s participation in the joint ICGI/CoNLL/LLL-2000 conference was fruitful, since he could share scientific views with his colleagues in the ICGI community, while managing to join in the work of the CoNLL community. This enabled him to meet with people involved in the TMR project. A co-authored paper (co-signed with C. De la Higuera) was presented at the ICGI conference. Some contacts were also established with H. Fernau from the University of Tübingen, Computer Science department. Franck and the latter were invited to write a joint conference report, which they did.

Since September, three collaborations have been made possible:

- As a result of the relationship built in Lisbon, Franck held a lecture at the Computer Science department of the University of Tübingen. The report of the ICGI/CoNLL/LLL conference was finished at this time.
- A seminar was also given at the France Telecom Research Lab and the automata
produced by MDI were plugged into their speech recognition system. Although the preliminary results were quite promising, the automata did not greatly affect the overall quality of the system. However, they dramatically sped up the response time of the system. Some trivial improvements of the plugging operation are currently being developed at the lab site.

- Some special links were made with the Geneva TMR team, and more specifically with Alexander Clark (the other new person in the project). Franck spent a week in Geneva in March 2000 in order to intensify this collaboration (see Alexander’s section).

Franck Thollard worked in two different directions:

2. Improving the probabilistic grammatical inference algorithms.

Since point 1 uses point 2 it was quite normal to work on the two fields. The preliminary results of the first item encourage research in this field. The second point aims at improving the first one. A paper was submitted for each of these topics, but the notifications of acceptance have not yet arrived.

### 6.7.2 Alexander Clark, PhD. Student

This section presents Alexander Clark’s participation in the project at Tübingen since he joined it on October 1, 2000. He is employed part-time at Tübingen, with another appointment at the TMR partner at ISSCO. The part-time appointment at Tübingen is in support of Clark’s scientific collaboration with Franck Thollard.

The topics of the collaboration are:

- Conditional Maximum likelihood training of transducers for the shared task.
- Formal aspects of PAC-learnability of finite state automata.
- Methods for introducing lexical probabilities into Franck’s chunking models.
- Examining a large bilingual corpus of patent abstracts from WIPO (World Intellectual Property Organisation) with regard to automatic extraction of terminology.

Franck visited ISSCO from 19 to 22 March.
6.7.3 Hervé Déjean, Postdoc

The participation of Hervé Déjean is presented here. Hervé's involvements with LCG will end 31 March 2001 (for a position at Xerox, Grenoble, LCG's industrial partner).

Hervé Déjean's project activities (April 2000 - March 2001) can be summarised as follows:

- **Shared task CoNLL-2000**: ALLiS was successfully applied on 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.

- The majority of the time was spent to introduce a hierarchical representation of the data, more adequate to the linguistic representation of the problem. For this purpose, a new version of ALLiS was written, using XML formalism. Integrating hierarchical data would ease learning of full-NP parsing.

- **Shared task CoNLL-2001**: In collaboration with Erik Tjong Kim Sang (Antwerp University), they prepared the next CoNLL-2001 shared task: Clauing, namely a segmentation into clauses. Clauing can be viewed as an intermediary step between chunking and full-NP parsing. The new version of ALLiS was also applied on this new data.

Hervé presented the following talks: **ALLiS: a Symbolic Machine Learning System For Natural Language Learning** (CoNLL-2000, September 2000), **Machine Learning and Corpus Linguistics** (University of Caen, October 2000), **Learning Set of Rules for Partial Parsing** (XRCE, December 2000)

His recent publications are: Déjean (2000a) Déjean (2000b) One publication is submitted to an IJCAI workshop, concerning unsupervised learning.

6.7.4 Dale Gerdemann, coordinator

Dr. Dale Gerdemann (Akademischer Rat, local coordinator) along with Franck Thollard and Hervé Déjean taught a course on finite state methods and machine learning for natural language processing (see below: "Training activities"). He has worked together with Franck Thollard on algorithms for transducer learning. He has also continued collaboration with Gertjan van Noord of Groningen University, with whom he has two recent papers: "Finite State Transducers with Predicates and Identity" (presented at Computational Linguistics in the Netherlands) and "Approximation and Exactness in Finite State Optimality Theory" (invited talk at SIGPHON 2000, COLING).
6.7.5 Related Tübingen Researchers

Prof. Dr. Erhard Hinrichs and Sandra Kübler M.A. are working on the application of symbolic machine learning techniques to parsing. They are using a similarity-based approach to extend partial chunk parses into complete tree structures, including function-argument structure.

Their related publications are:

- (Kübler & Hinrichs 2000)

Nathan Valette (PhD student) participated in the course taught by Dale Gerdemann. For the course he developed a monotonic second order logic interface for the Finite State Utilities of Gertjan van Noord. He was invited by Gertjan van Noord in Groningen to present his work. He is also working on HPSG, and presented a paper at the HPSG workshop in Tübingen.

Klaus Hoerman (M.A. student) is developing a system that is able to learn a Constraint Grammar for morphological disambiguation, using the morphological analyser Morphy, which returns a set of (almost) all the possible morphological tags for each word. The program is based on the MuTBl system which interprets a number of templates specifying the possible Constraint Grammar rules, and consequently learns a set of rules.

Tylman Ule (staff member), in the framework of the DEREKO (Deutsches ReferenzKorpus) project, uses Machine Learning techniques for tagging. For this, he built up training data for German and trained several taggers. He plans to learn morphology in the near future.

6.7.6 Training activities

Dale Gerdemann, Franck Thollard and Hervé Déjean have prepared and delivered a course on Machine Learning and Finite State Automata here in Tübingen during the winter semester of 2000.

Franck Thollard and Hervé Déjean attended a Tutorial (EAIA'00, Lisbon, September 2000) on the topic: Information Exploration and Learning Text.

References


A Shared Task Definition, Text Chunking

Text chunking consists of dividing a text in syntactically correlated parts of words. For example, the sentence

He reckons the current account deficit will narrow to only #1.8 billion in September.

can be divided as follows:

[NP He] [VP reckons] [NP the current account deficit] [VP will narrow] [PP to] [NP only #1.8 billion] [PP in] [NP September].

Text chunking is an intermediate step towards full parsing. It was the shared task for CoNLL-2000. Training and test data for this task is available. This data consists of the same partitions of the Wall Street Journal corpus (WSJ) as the widely used data for noun phrase chunking: sections 15-18 as training data (211727 tokens) and section 20 as test data (47377 tokens). The annotation of the data has been derived from the WSJ corpus by a program written by Sabine Buchholz from Tilburg University, The Netherlands.

The goal of this task is to come forward with machine learning methods which after a training phase can recognize the chunk segmentation of the test data as well as possible. The training data can be used for training the text chunker. The chunkers will be evaluated with the F rate, which is a combination of the precision and recall rates: $F = 2\times \text{precision}\times\text{recall} / (\text{recall} + \text{precision})$ [Rij79]. The precision and recall numbers will be computed over all types of chunks.
Background Information

In 1991, Steven Abney proposed to approach parsing by starting with finding correlated chunks of words [Abn91]. Lance Ramshaw and Mitch Marcus have approached chunking by using a machine learning method [RM95]. Their work has inspired many others to study the application of learning methods to noun phrase chunking. Other chunk types have not received the same attention as NP chunks. The most complete work is [BVD99] which presents results for NP, VP, PP, ADJP and ADVP chunks. [Vee99] works with NP, VP and PP chunks. [RM95] have recognized arbitrary chunks but classified every non-NP chunk as VP chunk. [Rat98] has recognized arbitrary chunks as part of a parsing task but did not report on the chunking performance.

Software and Data

The train and test data consist of three columns separated by spaces. Each word has been put on a separate line and there is an empty line after each sentence. The first column contains the current word, the second its part-of-speech tag as derived by the Brill tagger and the third its chunk tag as derived from the WSJ corpus. The chunk tags contain the name of the chunk type, for example I-NP for noun phrase words and I-VP for verb phrase words. Most chunk types have two types of chunk tags, B-CHUNK for the first word of the chunk and I-CHUNK for each other word in the chunk. Here is an example of the file format:

```
He       PRP     B-NP
reckons  VBZ     B-VP
the      DT      B-NP
current  JJ      I-NP
account  NN      I-NP
deficit  NN      I-NP
will     MD      B-VP
narrow   VB      I-VP
to       TO      B-PP
only     RB      B-NP
#        #       I-NP
1.8      CD      I-NP
billion  CD      I-NP
in       IN      B-PP
September NNP    B-NP
.
.
0
```

The O chunk tag is used for tokens which are not part of any chunk. Instead of using the part-of-speech tags of the WSJ corpus, the data set used tags generated by the Brill tagger. The performance with the corpus tags will be better but it will be unrealistic since for novel text no perfect part-of-speech tags will be available.
Appendix References

This reference section contains two parts: first the papers from the shared task session at CoNLL-2000 and then the other related publications.
CoNLL-2000 Shared Task Papers

General Articles