

Learning Computational Grammars
TMR Project Nr. ERBFMRXCT980237
4th Annual Report

John Nerbonne

December 19, 2002

Summary

This document describes the fourth and final year's progress of the TMR Project *Learning Computational Grammars* (LCG). Although different sites wound down their activities during the course of the year, LCG began the year with a 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 target their respective 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 fourth year was the continued open definition of tasks and the invitation to non-project teams to participation in common attack on important problems. For the 2001 *Conference on Natural Language Learning* (CoNLL), LCG created a task description and an attendant training and testing set, which was the focus of the public meeting, held in conjunction with 2001 meeting of the Association for Computational Linguistics (ACL) in Toulouse. Project coordinator John Nerbonne was invited to address CoNLL in one of its two long lectures. To cap the project, project participants James Hammerton, Miles Osborne, Susan Armstrong and Walter Daelemans edited a special issue of *Journal of Machine Learning Research* 2, March 2002. Several project papers were included.

Contents

| | | |
|----------|---|----------|
| 1 | Introduction | 5 |
| 2 | Training | 6 |
| 3 | Collaboration | 6 |
| 4 | Industrial Involvement | 8 |
| 5 | Site Reports | 8 |
| 5.1 | Antwerp | 8 |
| 5.1.1 | Erik Tjong Kim Sang | 9 |
| 5.1.2 | Training Activities | 9 |
| 5.1.3 | Collaborations | 9 |
| 5.1.4 | Walter Daelemans, Coordinator | 9 |
| 5.2 | SRI International, Cambridge, UK | 10 |
| 5.2.1 | Summary of LCG Project Activities at SRI | 10 |
| 5.2.2 | Anja Belz, Postdoctoral Researcher | 10 |
| 5.2.3 | Progress | 11 |
| 5.2.4 | Main Achievements, Highlights, Best Published Results, etc. | 11 |
| 5.2.5 | Future | 11 |
| 5.2.6 | David Milward, Coordinator | 11 |
| 5.2.7 | New Publications | 12 |
| 5.3 | Dublin | 12 |
| 5.3.1 | James Hammerton (Postdoc) | 12 |
| 5.3.2 | Dissemination | 14 |

| | | |
|-------|--|----|
| 5.3.3 | Collaborations | 14 |
| 5.3.4 | Ronan Reilly (Co-ordinator) | 15 |
| 5.3.5 | Training | 15 |
| 5.3.6 | Other language-related activities at computer science, UCD | 15 |
| 5.4 | ISSCO, Geneva | 15 |
| 5.4.1 | Training | 16 |
| 5.4.2 | Collaboration | 16 |
| 5.4.3 | Alexander Clark, Ph. D. Student/ Post-Doc | 16 |
| 5.4.4 | Franck Thollard | 17 |
| 5.4.5 | Other Reseachers | 18 |
| 5.4.6 | Susan Armstrong, Site Coordinator | 18 |
| 5.4.7 | Related Researchers at ISSCO | 19 |
| 5.5 | Xerox Research Centre Europe, Grenoble, France | 19 |
| 5.5.1 | Postdoc: Nicola Cancedda | 19 |
| 5.5.2 | Training Activities | 20 |
| 5.5.3 | Industrial Involvement | 20 |
| 5.5.4 | Project Related Activities at XRCE | 20 |
| 5.5.5 | Local Project Coordinator: Eric Gaussier | 20 |
| 5.5.6 | Other Researchers | 20 |
| 5.6 | Groningen | 21 |
| 5.6.1 | Stasinos Konstantopoulos, PhD student | 21 |
| 5.6.2 | Susanne Schoof, PhD student | 22 |
| 5.6.3 | John Nerbonne, Coordinator | 22 |

| | | |
|-------|---|----|
| 5.6.4 | Related Groningen researchers | 23 |
| 5.6.5 | Training activities | 23 |
| 5.6.6 | Industrial Involvement | 24 |
| 5.6.7 | Collaboration | 24 |
| 5.7 | Tübingen | 24 |
| 5.7.1 | Training Activities at Tübingen | 25 |
| 5.7.2 | Industrial Involvement | 25 |
| 5.7.3 | Collaborations | 25 |
| 5.7.4 | Researchers | 26 |
| 5.7.5 | Franck Thollard | 26 |
| 5.7.6 | Invited researcher: Yuval Krymolowski from Israel | 27 |
| 5.7.7 | Invited researcher: Wouter Jansen from Groningen | 28 |
| 5.7.8 | Dale Gerdemann, Site Coordinator | 29 |
| 5.7.9 | Related Tübingen Researchers | 29 |

1 Introduction

The final year of the “Learning Computational Grammars” project was marked by transitions. The project itself shifted into a mode in which emphasis rested on completing project plans and preparing reports, the most important of which are the report invited by the most important conference on applying machine learning techniques to natural language, the Conference on Natural Language Learning (CoNLL, 2001) (Nerbonne, Belz, Cancedda, Déjean, Hammerton, Koeling, Konstantopoulos, Osborne, Thollard & Tjong Kim Sang 2001), and the special issue of the *Journal of Machine Learning Research*, which was devoted to the task of the project, i.e., shallow parsing (structure recognition) in texts (Hammerton, Osborne, Armstrong & Daelemans 2002). Four of the seven papers in that special issue arose in the context of the LCG project:

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 Dejean (LCG Tübingen, then Xerox Grenoble)

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

Different groups finished at different times, and postdocs naturally needed to spend time securing follow-up positions. Three of the postdocs left to assume posts at LCG partners, viz., Hervé Déjean, who moved from Tübingen to Xerox, Grenoble; James Hammerton, who moved from Dublin to Groningen; and Franck Thollard, who moved from Tübingen to Geneva. Several other young researchers continued at the institutions where they held LCG positions, incl. Alexander Clark (Geneva), Erik Tjong Kim Sang (Antwerp), and Nicola Cancedda (Xerox, Grenoble); and others continuing in their host country at other institutions (Anja Belz and Rob Koeling, both formerly at SRI, Cambridge, now both in Brighton); and two predocs are completing their PhD dissertations (Susanne Schoof and Stasinou Konstantopoulos, both at Groningen). The LCG project has thus also contributed to the more permanent cross-fertilization of European research.

A final LCG meeting was held in Toulouse, in conjunction with the international meeting of the Association for Computational Linguistics in July 2001 (ACL 2001). Ten of the individuals involved in LCG gave 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 et al. 2001).

2 Training

Erik Tjong Kim Sang took part of a tutorial on Automatic Summarization presented by Inderjeet Mani and Mark Maybury at CoNLL-2001. Nicola Cancedda attended the 13th European Summer School in Logic, Language and Information (ESSLLI) in Helsinki, Finland, from August 13 to August 25, 2001. Moreover, he actively participated in the life of the Xerox, Grenoble group by leading a reading group on Machine Learning, which involved the participation of about 15 members, and attending to a number of scientific presentations. Stasinou Konstantopoulos (Groningen) also attended the 13th ESSLLI (2001) where he has had a paper accepted for the student session. A postgraduate course on statistical methods for parsing was held at Tübingen university during the spring semester of 2001. The LCG researchers (namely Hervé Déjean, Franck Thollard and Yuval Krymolowski) attended this research meetings. Franck Thollard taught 4 hours of these meeting sessions while Yuval Krymolowski provided 8 hours; he gave a 4 hour presentation of his system (see below for more information on his system) and another 4 hour practical talk. Wouter Jansen, the final LCG researcher at Tübingen, was trained in finite-state methods and Optimality Theory by Dale Gerdemann; Jansen in turn trained Katja Jasinskaja, a PhD student in the Department of Linguistics, in the design of psycholinguistic experiments.

Stasinou Konstantopoulos (Groningen) also assisted Gosse Bouma by giving tutorials for a course in Natural Language Processing. (March - June 2002), see [GosseNLP{http://www.let.rug.nl/~gosse/nlp1}](http://www.let.rug.nl/~gosse/nlp1) for more details He has also attended various talks on subjects of general interest (e.g. the weekly linguistics Colloquium in Groningen) and has also presented his work in the Groningen Linguistics Colloquium (May 2002). Susanne Schoof (Groningen) assisted Dr. Leonie Bosveld in an introductory course on information science.

Franck Thollard, postdoc at Tübingen and Geneva, supervised Master's students at the University of St. Etienne.

The LCG senior researchers at universities are all involved in training activities, and the LCG research found its way into these in the normal way. Prof. Dr. Erhard Hinrichs (Tübingen) also organized a workshop on “Machine Learning Approaches in Computational Linguistics” which has since been held at the ESSLLI 2002 summer school in Trento.

ISSCO also hosted several young researchers during this period.

3 Collaboration

The major collaboration of the final LCG year was editing of a special issue of the *Journal of Machine Learning Research* on “Machine Learning Approaches to Shallow Parsing”. This is now online at <http://www.jmlr.org/>. This involved James Hammerton (Dublin) Susan Armstrong (Geneva), Walter Daelemans (Antwerp) and Miles Osborne (formerly at LCG Groningen).

John Nerbonne, Hervé Déjean (Tübingen) and Erik Tjong Kim Sang (Antwerp) jointly gave an invited talk on the *Learning Computational Grammars's* project at CoNLL-2001 in Toulouse (Nerbonne et al. 2001). Nerbonne spent the Winter semester (2001-2002) as guest professor at Stuttgart, where his course on machine learning applied to natural language was also attended by students from LCG partner Tübingen.

In this final project year, Erik Tjong Kim Sang (Antwerp) confirmed his (unofficial) role as “hub” of LCG collaboration efforts by collaborating with Hervé Déjean (Tübingen) in the organization of the shared task of the workshop Computational Natural Language Learning (CoNLL-2001). He has continued with regularly paying a visit to Antwerp’s sister group in Tilburg, The Netherlands. Dr. Tjong Kim Sang also co-authored a paper on a hybrid architecture, involving a combination of the (neurally inspired) self-organizing map and the memory-based learning system, with James Hammerton (Dublin). The new cooperative effort was based on their earlier collaboration within the LCG project and resulted in a presentation at the CoNLL 2001 workshop. This system achieves performance close to that of MBL but using only a fraction of the number of comparisons.

There was also extensive collaboration between Alexander Clark in Geneva and Frank Thollard at the Tuebingen site and at Geneva. Alexander Clark also gave a talk at Xerox Research Centre Europe on a method for learning Finite-State Transducers. Walter Daelemans of Antwerp was the external examiner for Alexander Clark’s D. Phil. Thesis at the University of Sussex.

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 (Groningen) FSA Utilities¹. The talk was given in the context of the weekly colloquium of CLCG².

Frank Thollard (Tübingen) visited Groningen in September 2001 to deliver a talk *On the understanding of the 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.

Combining inter-LCG collaboration with industrial involvement, Rob Koeling (SRI Cambridge) 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 “Industrial Involvement”).

Further collaboration between Groningen and Tübingen was achieved through the work of Groningen’s Wouter Jansen who became the final LCG researcher in Tübingen.

¹<http://odur.let.rug.nl/~vannoord/Fsa/>

²See <http://www.let.rug.nl/clcg/> about CLCG and the colloquium

4 Industrial Involvement

XRCE is an industrial research center. Research conducted at XRCE, including that issuing from the LCG project, directly or indirectly impacts a number of the products and services in the offer of Xerox, especially in the field of document and knowledge management.

SRI Cambridge is likewise 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.

Combining inter-LCG collaboration with industrial involvement, Rob Koeling (SRI Cambridge) visited Groningen between Sept. and Dec. 2001, during which time he worked together with Gosse Bouma (Alfa-Informatica), Tanja Gaustad (Alfa-Informatica, ALP project) and John Nerbonne on an email classification project³ contracted by the Groningen-based company, BSC⁴.

The Tübingen researcher Hervé Déjean developed close contacts with Xerox XRCE. Through these contacts, Dr. Déjean was invited to join this French research Team Xerox.

The other full-time Tübingen researcher (namely Franck Thollard) has some contact with France Télécom R&D⁵ and some experiments were made to plug his model in the industrial partner spontaneous speech recognizer. His model has been shown to be nearly as accurate as theirs but much more faster.

5 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, 2001 – March 31, 2002.

5.1 Antwerp

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

³<http://vili.let.rug.nl/zonnet/>

⁴<http://www.bsc.nl/>

⁵France Télécom R&D is the French Télécom research lab. Franck Thollard wrote his PhD dissertation under a grant provided by France Télécom R&D.

5.1.1 Erik Tjong Kim Sang

In the fourth project year (April 1, 2001 - March 31, 2002), Erik has worked on clause identification and full parsing. He co-organized the shared task of CoNLL-2001 and submitted papers to CoNLL-2001, CLIN-2000 and the Journal of Machine Learning Research. Erik's position at the LCG project ended at July 15, 2001.

The clause identification work has been done in the framework of the CoNLL-2001 shared task. Together with Hervé Déjean from Tübingen, Erik developed the shared task and evaluated the results that were submitted. He participated with a memory-based shallow parser in this shared task. The system obtained the third-best results from six participants. This work was published in a shared task paper at CoNLL-2001. Erik also published an introduction paper to the shared task in the proceedings of this conference (joint work with Hervé Déjean).

The parsing work was a follow-up of work performed in the previous year. A paper describing the memory-based chunk parser has been accepted for the proceedings of CLIN-2000 (to appear in November 2001)(MARIETTE). Erik's joint work with James Hammerton of Dublin has led to a joint paper presented at CoNLL-2001. Apart from this, he has also been active as a co-author of an overview paper of the LCG project which was presented at the same conference. Erik has also submitted an overview paper of his work on the project to the Journal of Machine Learning Research which will publish a special issue on Machine Learning Approaches to Shallow Parsing in 2002. In this paper he presented some of his most recent work: using feature selection methods for finding optimal feature sets for memory-based learners.

5.1.2 Training Activities

In the fourth project year, Erik took part of a tutorial on Automatic Summarization presented by Inderjeet Mani and Mark Maybury at CoNLL-2001.

5.1.3 Collaborations

In this final project year, Erik has collaborated with Hervé Déjean (Tübingen) in the organization of the shared task of the workshop Computational Natural Language Learning (CoNLL-2001). He has continued with regularly paying a visit to Antwerp's sister group in Tilburg, The Netherlands.

5.1.4 Walter Daelemans, Coordinator

Walter Daelemans continued to act as site coordinator during the final year of the project. Earlier work on shallow parsing led to a journal publication on the usefulness of shallow

parsing for Question Answering, and the co-editing of a special issue on shallow parsing for JMLR.

CNTS entered a National Science Foundation-sponsored research community on Machine Learning and Datamining (responsible for Textmining), and started two new funded research projects, one on extraction of ontologies from text using unsupervised learning (dr. M. Reinberger), and one on learning lexical semantics from text using clustering techniques (B. Decadt). Together with other groups at the University of Antwerp, funding was obtained for a parallel computer cluster which CNTS will use to experiment with parallel processing for computationally intensive machine learning of language experiments. Currently, 2 postdocs and 4 predocs are full-time employed working on subjects directly related to Machine Learning of Natural Language. The expertise developed in machine learning of language also continues to have an impact on the research of the people within CNTS working on computational psycholinguistics (building models of human language acquisition and processing). Within this group, 2 pre-docs use machine learning methods in their research.

5.2 SRI International, Cambridge, UK

During the reporting period (May 2001 – August 2001) SRI employed one postdoctoral researcher. This section contains a summary of general project activities at SRI, and a more detailed report on the activities of the postdoctoral researcher.

5.2.1 Summary of LCG Project Activities at SRI

LCG work at SRI continued until the end of August 2001, when the remaining postdoctoral researcher left to take up a new position. During the four months prior to her departure, work focussed on the completion of ongoing research, and the publication of papers reporting the results.

5.2.2 Anja Belz, Postdoctoral Researcher

During the reporting period, Anja Belz continued to work with probabilistic context-free grammars (PCFGs) that incorporate *Local Structural Context* (LSC). The main focus of her final research project for 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 LCG project tasks. In this applica-

tion, base grammars incorporating local structural context are derived in a preliminary step from the Wall Street Journal Corpus WSJC. Then, the space of partitions of the base grammar’s set of nonterminals is searched for a partition that maximises parsing performance and minimises grammar size. The net result of this approach is an efficient, task-specific grammar that incorporates just the structural context that is useful for the given task.

5.2.3 Progress

A large number of experiments were carried out for Partition Tree Search Learning of PCFGs, involving a range of partial and complete learning tasks. Results for the complete parsing task are better than the best published results for WSJC data and nonlexicalised parsers. Results for the partial parsing tasks are also better than the best published results (where these exist) for nonlexicalised parsers, and are close to current best results achieved by lexicalised systems.

5.2.4 Main Achievements, Highlights, Best Published Results, etc.

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).

5.2.5 Future

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.

5.2.6 David Milward, Coordinator

David Milward has continued to supervise Anja Belz. He has looked into the use of the maximum entropy approach to noun group chunking within the SRI Highlight Information Extraction engine.

5.2.7 New Publications

1. Anja Belz (2001) *Learning Local Structural Context Grammars for Different Parsing Tasks by Partition-Tree Search*. Also submitted to Journal of Machine Learning Research for *Special Issue on Machine Learning Approaches to Shallow Parsing*.
2. Anja Belz (forthcoming) *Learning Broad-Coverage Grammars for Different Parsing Tasks from Parsed Corpora: Report on Research for TMR Project 'Learning Computational Grammars'*.

5.3 Dublin

This report summarise the progress made on this project since April 2001. Note that James Hammerton left the project at the end of 2001 to start a new postdoctoral research post at the University of Groningen, and that Ronan Reilly left UCD, also at the end of 2001, to take up the post of Head of the Computer Science Department, at the National University of Ireland, Maynooth.

5.3.1 James Hammerton (Postdoc)

Long Short-Term Memory (LSTM), At the time of the previous report, James was investigating the use of LSTM networks for the noun-phrase bracketing task and in particular investigating whether new ways of representing the problem to the networks would improve the speed at which the task can be learned.

Significant progress has been made since then. A new approach to the task was found to involve quicker training than the earlier approach did. Each sentence is presented, word by word, to the network in two passes:

- On the first pass, the word and its corresponding POS tag are presented as input, but the network is trained not to produce any output. The first pass is used to accumulate information for disambiguation in the second pass.
- On the second pass the sentence is presented word by word again, and this time as each word is presented the network is trained to output the number of noun-phrases beginning or ending on that word.
- A unit in the input layer is reserved for indicating which pass through the sentence is currently being processed.

Unlike the earlier approach the output representation allows sentences of arbitrary length to be processed. The only limitation imposed by this approach is a fixed upper limit on the number of noun-phrases starting or ending at each word.

| Hidden Layer | Sentences used | Train fscore | Test fscore |
|---------------|------------------|--------------|-------------|
| 12×4 | Length < than 10 | 98.19 | 70.14 |
| 6×4 | Length < than 20 | 75.44 | 72.01 |
| 8×4 | Length < than 20 | 77.22 | 71.25 |

Table 1: Selection of results for noun-phrase bracketing on sentences from sections 15 to 18 (train) and section 20(test) of the Wall Street Journal corpus.

| 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 2: Selection of results for clause identification on sentences from sections 15 to 18 (train) and section 20(development) and section 21(test) of the Wall Street Journal corpus. The networks were trained on the first 1000, 2000 and 3000 sentences of sections 15 to 18 of the WSJ corpus respectively.

Training was much quicker with this approach reflecting lower burdens placed on the network than with the earlier approach. As a result, LSTM was applied to bracketing the noun-phrases for all sentences of length < 10, yielding a best noun-phrase fscore of 70.14 on the testing set, using a network with a hidden layer of 12×4 cells.

To extend this work, the new approach was also applied to all of the sentences of length < 20 from the training and testing sets, yielding a best fscore of 72.04 on the testing set, using a network with a hidden layer of 6×4 cells. See Table 1 for details.

Clause splitting with LSTM The improved training times enabled James to put in an entry for the CoNLL 2001 shared task. Part 3 of the task involves finding all the clauses in a sentence, and this is the part which James trained his networks to do. They were trained on the first 1000 and 2000 of the sentences, yielding clause fscores of 45.61 and 50.42 respectively on the testing set. Whilst this is above the baseline it was worse than the other entries in the CoNLL workshop which were in the range 62.77 to 78.63.

However this may in part be due to the use of only part of the training set which totals 8936 sentences and lack of time to get better performance on the training set. Unfortunately, even with the faster training, it still takes a long time to train the networks. Since the CoNLL workshop LSTM has been applied to the first 3000 sentences on this task yielding an fscore of 52.14. See Table 2 for details.

Forget gates and Peephole connections There have been some recent additions to LSTM that it is claimed improve LSTM’s performance. Forget Gates link directly to the cells in LSTM, and learn to modify the values stored in the cells. Peephole connections

connect the cells to the gates, enabling the cell contents to influence the behaviour of the gates directly.

In order to see whether it obtains better performance, James has added to the LSTM code so that it supports forget gates and peephole connections. At the present time, the code for forget gates seems to work, but the code for peephole connections is still being worked on. It is hoped that James will be able to finish off this work in his next post.

5.3.2 Dissemination

Paper for JMLR Special Issue 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 revised and resubmitted as a regular JMLR paper. It is hoped the revised paper, which will include work employing forget gates and peephole connections as well as hopefully using all training data, can be produced in James' next post.

- James attended the “Computational Natural Language Learning Workshop (CoNLL-2001)” held in Toulouse, France on the 6 and 7th of July, where he presented 2 papers, 1 in the main session (see below) and 1 for the shared task (see above).

5.3.3 Collaborations

James was involved in the following collaborations:

- Co-editing 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 online at “<http://www.jmlr.org/>”. This has given James valuable experience of the process of editing a journal.
- A paper on a Hybrid SOM/MBL system, co-authored with Erik Tjong Kim Sang, produced as a result of James' earlier collaboration with the LCG group in Antwerp, was presented at the CoNLL 2001 workshop. This system achieves performance close to that of MBL but using only a fraction of the number of comparisons.

James has since performed experiments looking at the impact of varying network size and the number of winners chosen during testing on the performance of the system. He has written, but not yet tested, code for taking into account the distribution of points in each cluster when selecting the winning cluster. He hopes to be able to finish this work off and produce a followup paper in his next post.

5.3.4 Ronan Reilly (Co-ordinator)

Within the Cognitive and Computational Neuroscience Centre (CCNC), Ronan has been involved with another strand of language related research activity which has concentrated on the neural basis for language learning and language processing.

Much of this work has focussed on a theory of cortical development proposed by Reilly (2001) called Cortical Software Re-Use (CSRU), which not only can be applied to language development but also to cognitive development in general. Over the last year the main research focus has been on developing a mathematical basis for CSRU (Reilly & Mackey 2001).

5.3.5 Training

Ronan Reilly taught various courses to undergraduates and postgraduate at UCD, including courses for the MA/MSc in Cognitive Science.

5.3.6 Other language-related activities at computer science, UCD

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.

The principal interest of this group is to develop robust and motivated approaches to speech recognition and synthesis. Research focuses on the linguistic and cognitive motivations underpinning speech. The group is concerned with developing computational linguistic models, and investigating cognitive and psycholinguistic issues. In developing these areas, knowledge of structures and linguistic expectations (such as phonotactics), temporal constraints, and the human influence on speech (the embodiment of speech) are incorporated.

Relevant publications include (Ashby, Carson-Berndsen & Joue 2001, Bohan, Creedon, Carson-Berndsen & Cummins 2001, Carson-Berndsen, Joue & Walsh 2001, Carson-Berndsen & Walsh 2001, Carson-Berndsen & Gibbon 2001, Cummins & Deb. 2001, Cummins 2001*a*, Cummins 2001*b*, Joue & Carson-Berndsen 2001, Nenova, Joue, Reilly & Carson-Berndsen 2001, Nenova & Reilly 2001).

5.4 ISSCO, Geneva

This report covers the period April 1 2001 to March 31 2002.

This site employed one PhD student throughout the year, Alexander Clark, and one post doc for 5 months Franck Thollard, . 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.

5.4.1 Training

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.

ISSCO also hosted several young researchers during this period.

5.4.2 Collaboration

There was extensive collaboration between Alexander Clark at this site, and Franck Thollard at the Tuebingen site and at Geneva. Alexander Clark gave a talk at Xerox Research Centre Europe on a method for learning Finite-State Transducers. Walter Daelemans of Antwerp was the external examiner for Alexander Clark's D. Phil. Thesis at the University of Sussex.

5.4.3 Alexander Clark, Ph. D. Student/ Post-Doc

Alexander Clark has continued his researches into the field of unsupervised language learning. During the period under review this has resulted in three publications.

- A paper entitled *Learning Morphology with Pair Hidden Markov Models*, which was presented at the Student Workshop at the 39th Annual Meeting of the Association for Computational Linguistics, in Toulouse in July. This paper addressed the Supervised learning of Morphology using a statistical model.
- A paper entitled *Unsupervised Induction of Stochastic Context Free Grammars with Distributional Clustering*, presented at the Workshop on Computational Natural Language Learning (CoNLL 2001). This presented an unsupervised algorithm for inducing context free grammars.
- A paper entitled *Partially Supervised Learning of Morphology with Stochastic Transducers*, presented at the Natural Language Processing Pacific Rim Symposium, in Tokyo. This paper showed how supervised learning, can be converted to unsupervised learning in the field of morphology.

In addition he submitted his D. Phil. thesis entitled *Unsupervised Language Acquisition: Theory and Practice*, at the University of Sussex, and successfully defended it. This thesis addressed the relevance of recent advances in Machine Learning to the so-called Argument of the Poverty of the Stimulus, that claims that the information available to the infant child when he learns his native language is too impoverished to allow learning to proceed.

During the period in question, Alexander Clark of Geneva collaborated with Franck Thollard (of Tübingen and later Geneva) on a number of different areas, primarily concerned with grammatical inference, and its application to natural language learning.

- The incorporation of lexical information into grammatical inference: since the algorithms used tend to be feasible only with small vocabulary sizes, it is necessary to use some sort of clustering algorithm to create a small number of classes of words. The application here is in Language Modelling.
- The use of grammatical inference in the shared task, again using lexical information to improve the results. Here the approach is to identify particular areas where lexical information will improve the results, and then to identify particular words that will help.
- Some formal proofs of the convergence of the algorithms used in two frameworks: the identification in the limit with probability one framework, and the PAC-learnability framework.

Some of these are currently being prepared for publication.

Other than the papers mentioned above, Alexander Clark also gave a number of invited talks:

- *Learning Finite State Transducers with Pair Hidden Markov Models*, at Xerox Research Center Europe, Grenoble on May 16th 2001.
- *Learning Finite State Transducers with Pair Hidden Markov Models*, at IDIAP, Martigny, on June 12th 2001.
- *Formal methods in Natural Language Processing*, at EURISE, University of St. Etienne, France, on January 24th 2002.

5.4.4 Franck Thollard

Franck Thollard joined ISSCO in September 2001 and was employed there until February 2002. He first joined the Tübingen site and then went to Geneva in order to work more closely 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 LCG researchers (mainly 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 improves the chunker so that it provided up to 90% of good answers. This work has since project end been presented at the International Conference

on Grammatical Inference (ICGI 2002). In addition, Franck Thollard and Alexander Clark have worked on a more theoretical paper. The two works (NP-Chunker and theoretical one) were published in the proceedings of the French Conference on Machine Learning (2002). The international version of the theoretical work is under submission at the Algorithmic Learning Theory (ALT) 2002 conference.

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 also collaborates 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.

5.4.5 Other 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 resources 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.

5.4.6 Susan Armstrong, Site Coordinator

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.

5.4.7 Related Researchers at ISSCO

Pierrette Bouillon continued to work on the task of automatic acquisition of lexical semantics using a symbolic learning algorithm.

5.5 Xerox Research Centre Europe, Grenoble, France

XRCE participates in the LCG project with one postdoc. This section briefly describes his work and that of other related researchers.

5.5.1 Postdoc: Nicola Cancedda

The research focus shifted—in the last project year—from Explanation Based Learning for grammar specialisation and adaptation to probabilistic models for syntactic structural disambiguation, with a special emphasis on attachment disambiguation problems. Research related to the LCG project conducted in the period under consideration includes:

- The development of a probabilistic dependency language model. This model estimates attachment probabilities for prepositional, adjectival and adverbial phrases based on lexical, lexical semantic and syntactic information as obtained from a shallow parser. The proposed model generalises over several others previously developed to the same purpose.
- The development of an unsupervised training procedure for estimating the probabilities in the model above. The procedure relies on *natural supervision*, i.e. the observation that in some special configuration the attachment decision is not ambiguous. Such special configurations are used to gather statistics which, once smoothed, provide estimates for model probabilities, under the assumption that the relevant probability distributions are the same in ambiguous and non-ambiguous configurations. Interesting results were achieved by the model, regardless of this underlying assumption not being in general satisfied.
- The design and implementation of a bare-bone dependency parser, following the approach of (Eisner 2000). The dependency parser relies on the probabilistic model for producing the most likely set of dependencies for a given sentence. The parser is bare-bones in the sense that the dependencies themselves are not labeled.
- A series of experiments conducted to validate the model above, using training sets of up to 300,000 sentences from the French newspaper *Le Monde*. The probabilistic model, the estimation technique and the experiments are described in (Gaussier & Cancedda 2001a, Gaussier & Cancedda 2001b).

Nicola Cancedda also contributed to research on kernel methods for text categorisation, clustering and filtering in the context of the EC project KerMIT.

5.5.2 Training Activities

Nicola Cancedda attended the 13th European Summer School in Logic, Language and Information in Helsinki, Finland, from August 13 to August 25, 2001. Moreover, he actively contributed to training by leading a reading group on Machine Learning, which involved the participation of about 15 members, and attending to a number of scientific presentations.

5.5.3 Industrial Involvement

XRCE is an industrial research center. Research conducted at XRCE, including that issuing from the LCG project, directly or indirectly impacts a number of the products and services in the offer of Xerox, especially in the field of document and knowledge management.

5.5.4 Project Related Activities at XRCE

Several other projects concerned with the use of machine learning techniques for text processing were active at XRCE in the period under consideration. They include:

- Probabilistic models for hierarchical document clustering and categorisation;
- Corpus-based multilingual thesaurus enrichment;
- Inductive learning for typed entity recognition in information extraction applications;
- Kernel-based document categorisation, clustering and filtering;
- Language modeling for post-OCR correction and accent recovery;

5.5.5 Local Project Coordinator: Eric Gaussier

Besides supervising the work of the XRCE postdoc, Eric Gaussier coordinated the activities of the *Machine Learning for Natural Language Processing* group at XRCE. He actively participated in all the ongoing projects, ranging from probabilistic models for hierarchical document categorisation and clustering (Gaussier, Goutte, Papat & Chen 2002), to kernel-based methods, machine-learning methods for information extraction and post-OCR correction. All these projects resulted in a number of publications on refereed journals and of presentations at international conferences.

5.5.6 Other Researchers

Lemine Abdellahi joined XRCE mid July 2001. He worked on the design of statistical language models, with a special concern for the problem of correcting the output of an Optical

Character Recognition system. Moreover, he applied the same language modeling technique to the problem of recovering accents in written text, a particularly relevant problem in the case of the French language.

Hervé Déjean, formerly LCG postdoc at Tuebingen University, joined XRCE mid April 2001. His work concerned the creation and enrichment of a bilingual lexicon in the medical domain by means of the combined use of a specialised thesaurus, a bilingual corpus and a general-purpose bilingual lexicon. This work was mostly carried out in the context of the Muchmore joint EC-IST/NSF project. ALLiS, the rule induction tool developed at Tuebingen, is currently used in a French project, Biomire, in order to detect biological entities (genes, proteins) in biomedical documents. An article (Déjean 2002) appeared in *Journal of Machine Learning Research* describes this tool.

Moreover, Hervé Dejean coordinated the development of an interactive tool for annotating typed entities in document collections.

Cyril Goutte joined XRCE in October 2001. His work concerned kernel-based and probabilistic methods for document processing. In particular, he focused on the use of Support-Vector Machines for document categorisation in the context of the EC-IST project KerMIT, as well as on hierarchical probabilistic models for clustering and categorisation. He also worked on the connection between probabilistic methods and kernel-based discriminative methods using Fischer kernels, especially in relation with the joint use of annotated and unannotated data for training.

Jean-Michel Renders joined XRCE at the beginning of July, 2001. In the period under consideration his work was mainly concerned with kernel-based methods for text categorisation, clustering and filtering in the context of the KerMIT EC-IST project. His work concerned mostly the investigation of appropriate linguistic processing in conjunction with kernel-based methods as well as the design of linguistically motivated kernels.

5.6 Groningen

This site employs 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.

5.6.1 Stasinou Konstantopoulos, PhD student

Stasinou is using the Aleph Inductive Logic Programming System developed in Oxford (Srinivasan 2001) to induce Dutch Phonotactics from the Dutch section of the CELEX corpus. The results of this research have been presented in the student session of the 13th European Summer School on Language, Logic and Information (ESSLLI 2001) that took place in Helsinki in August 2001, and also published as part of the Proceedings of the student session (Konstantopoulos 2001, and also <http://www.helsinki.fi/esslli/> for more in-

formation on ESSLLI 2001). A journal paper version has been accepted for publication at the WEB-SLS on-line journal (<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 (Yet Another Prolog) equipped with an Message-Passing Interface (MPI) 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.

Non-LCG academic activities include attending 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.

5.6.2 Susanne Schoof, PhD student

Susanne spent her first year in Groningen (2000–2001) acquiring the necessary linguistic skills. She completed a project on verbal complementation, and during the past year her investigation was continued, now focussing on the behaviour of the reflexive pronoun in different structures. The predicted differences were found by means of corpus research, leading to a first formalisation in Head-Driven Phrase Structure Grammar (HPSG). As certain aspects shown by the data cannot easily be explained within the framework of HPSG alone, functionalist arguments are under investigation in order to provide deeper insight into the theoretical problem.

Susanne has given talks on her work on her syntactic analyses at a meeting of Dutch linguists (Katwijk, November 2001) and made poster presentations of the Behavioral and Cognitive Neurosciences (BCN) in Groningen (BCN poster day, February 2002) and Amsterdam (LATLING Conference, June 2001). Her submission to the HPSG conference (Seoul, August 2002) has also been accepted.

5.6.3 John Nerbonne, Coordinator

John Nerbonne has continued investigations into the application of unsupervised learning to the problem of dialect classification with (Heeringa, Nerbonne & Kleiweg 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 subject of language learning (Nerbonne 2002) and continued his work in phonological learning with *Learning Lexical Phonotactics with Simple Recurrent Network* (with Ivilin Stoianov, submitted to *Computer Speech and Language*, 2001.)

Finally, his review of Stefan Müller’s HPSG treatment of German (Müller 1999) was published in the *Journal of Germanic Linguistics* 13, 2001.

Nerbonne spent the Winter semester (2001-2002) as guest professor at Stuttgart (see section on “Collaboration”).

5.6.4 Related Groningen researchers

Gertjan van Noord is the project manager of the NWO PIONIER project *Algorithms for Linguistic Processing*. 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 training stochastic models) are tried for the ambiguity problem. See <http://www.let.rug.nl/vannoord/alp/> more information on ALP.

Rob Malouf joined the Groningen group in July 1999. During the last year he has been working on maximum entropy-based statistical natural language processing as part of the *Algorithms for Linguistic Processing* project (see above). 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 successfully defended his Groningen PhD thesis on *Dialogue-Based Disambiguation* in January 2002 (Koeling 2002). This involved applying maximum entropy learning to the disambiguation problems in a speech understanding system.

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 (in co-operation with the ALP project) have been presented in the NLP Pacific Rim Symposium (Mullen, Malouf & van Noord 2001).

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).

5.6.5 Training activities

Stasinos Konstantopoulos is assisting Gosse Bouma by giving tutorials for a course in Natural Language Processing (March - June 2002, see [GosseNLP{http://www.let.rug.nl/~gosse/nlp1}](http://www.let.rug.nl/~gosse/nlp1) for more details). He has 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 attended various talks on subjects of general interest in computational linguistics

(e.g., the weekly linguistics Colloquium in Groningen) and has also presented his work in the Groningen linguistics Colloquium (May 2002).

5.6.6 Industrial Involvement

Combining inter-LCG collaboration with industrial involvement, Rob Koeling (SRI Cambridge) visited Groningen between Sept. and Dec. 2001, during which time he worked together with Gosse Bouma (Alfa-Informatica), Tanja Gaustad (Alfa-Informatica, ALP project) and John Nerbonne on an email classification project⁶ contracted by the Groningen-based company, BSC⁷.

5.6.7 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). Nerbonne spent the Winter semester (2001-2002) as guest professor at Stuttgart, where his course on machine learning applied to natural language was also attended by students from LCG partner Tübingen.

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 FSA Utilities⁸. The talk was given in the context of the weekly colloquium of CLCG⁹.

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.

5.7 Tübingen

The following presents the activities of the Tübingen group in the LCG project. The training activity is first presented. The collaborations developed between Tübingen and other sites—namely the Geneva site and Groningen university—is described. Then the project researchers are presented along with their work related to the project.

⁶<http://vili.let.rug.nl/zonnet/>

⁷<http://www.bsc.nl/>

⁸<http://odur.let.rug.nl/vannoord/Fsa/>

⁹See <http://www.let.rug.nl/clcg/> about CLCG and the colloquium

5.7.1 Training Activities at Tübingen

A postgraduate course on statistical methods for parsing was held at Tübingen university during the spring semester of 2001. The LCG researchers (namely Hervé Déjean when he was still in the project, Franck Thollard and Yuval Krymolowski) attended this research meetings. Franck Thollard taught 4 hours of these meeting sessions while Yuval Krymolowski provided 8 hours; he gave a 4 hour presentation of his system (see below for more information on his system) and another 4 hour practical talk. Wouter Jansen, the final LCG researcher at Tübingen, was trained in finite-state methods and Optimality Theory by Dale Gerdemann. He also collaborated with Katja Jasinskaja, a PhD student in the Department of Linguistics. Wouter Jansen trained Katja Jasinskaja in the design of psycholinguistic experiments.

5.7.2 Industrial Involvement

The Tübingen researcher Hervé Déjean developed close contacts with Xerox XRCE. Through these contacts, Dr. Déjean was invited to join this French research Team Xerox.

The other full-time Tübingen researcher (namely Franck Thollard) has some contact with France Télécom R&D¹⁰ and some experiments were made to plug his model in the industrial partner spontaneous speech recognizer. His model has been shown to be nearly as accurate as theirs but much faster.

5.7.3 Collaborations

A strong collaboration was formed with the main ISSCO researcher (namely Alexander Clark) and some mutual reviewing of articles and writing was done. For this collaboration, Clark worked part time in Tübingen (1.10.2000—31.12.2001). In September 2001, Franck Thollard moved to Geneva to further this collaboration, and then moved back Tübingen for the final two months. He has thus taken great advantage of the mobility opportunity offered by the LCG project.

Moreover, another collaboration between Groningen and Tübingen started and Franck Thollard gave two talks in Groningen: a tutorial on grammatical inference with a particular emphasis on the probabilistic aspect and a talk more oriented on the Franck Thollard algorithms. Moreover these two sites aim at sharing some code. Further collaboration between Groningen and Tübingen was achieved through the work of Wouter Jansen who, coming from Groningen, became the final LCG researcher in Tübingen.

A collaboration with the other sites of the project led to a joint article at the CoNLL 2001 conference (Nerbonne et al. 2001)

¹⁰France Télécom R&D is the French Télécom research lab. Franck Thollard wrote his PhD dissertation under a grant provided by France Télécom R&D.

5.7.4 Researchers

Tübingen has had a main researcher (namely Franck Thollard) and an invited one, Yuval Krymolowski, from Israel. The final researcher was Wouter Jansen from Groningen, who worked briefly late in the project. The work of these researchers is presented in the following sections. Hervé Déjean left the project as he obtained a position at the Xerox European center. Thus the LCG project has been very successful in furthering the career of this young researcher.

This following sections present the research activity of the Tübingen LCG researchers. Franck Thollard's activity is presented first followed by the activities of Yuval Krymolowski and Wouter Jansen.

5.7.5 Franck Thollard

Franck Thollard worked in two directions:

- improving the probabilistic grammatical inference algorithm techniques themselves.
- applying state of the art probabilistic grammatical inference techniques to the project (on the NP chunking task).

The first item led to an article at the ICML conference¹¹ (Thollard 2001*a*). Many interesting contacts were made there. The paper provides a new probabilistic grammatical inference algorithm and some ensembles of methods, which out-perform the state of the art result of this kind of technique on a classic natural language modeling task.

For the second item, Franck Thollard presented the application of these new techniques to the NP-Chunking task at the CAP¹² conference (Thollard 2001*b*). The NP-Chunking problem was a shared task of the CoNLL 2000 conference. This task was built up by elements of the LCG project (mainly by Erik Tjong Kim Sang).

Taking advantage of the mobility opportunities provided by the LCG project Franck Thollard moved to Geneva for six months and moved back to Tübingen for the final two months of the project. Up to the end of the project, he worked mainly with Alexander Clark, the main Geneva researcher. Their collaboration led to an improvement of the work initiated by Franck Thollard on the chunking task. By the end of the project, they worked together to improve the system started by Franck Thollard at Tübingen: dealing with the CoNLL2000 shared task initiated by the LCG members. Their joint work raised the system performance up to the level of 90% accuracy.

¹¹ICML stands for International Colloquium on Machine Learning. It was held this year (2001) in Williams. The acceptance range of papers was 80 out of the 249 submitted.

¹²CAP stands for Conférence d'Apprentissage which is the main French conference on machine learning. The article was accepted as a "Long article". The selection rate for the "long article" was 16 accepted article out of 27 submitted.

Another domain on which they worked together was more theoretical. They worked on a formal learning definition. Some learnability results on a well known paradigm were proven and a new definition of what learning means was provided. This results where accepted at the French machine learning conference (namely CAP 2002) and are currently under submission at international conferences:

- International Conference on Grammatical Inference for the chunking system.
- Algorithmic Learning Theory for the theoretical results.

Franck Thollard also started a collaboration with some member of EURISE, the Saint-Etienne University (France) research team. This collaboration led to an article at the European Conference on Machine Learning.

Moreover, Franck Thollard and Robbert Prins from Groningen University are still collaborating in order to use Franck Thollard's models to speed up the Groningen University Dutch system.

As regard to the Training activities, Franck provided a talk at IDIAP (Martigny, Switzerland).

Thollard was asked to be a reviewer of the international journals JMLR (Journal of Machine Learning, 2 papers) and for the NLE (Natural Language Engineering).

5.7.6 Invited researcher: Yuval Krymolowski from Israel

Yuval is a PhD student at the department of computer science in Bar-Ilan university, Israel. His supervisor is Ido Dagan. Yuval visited Tübingen between May and August 2001.

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, Yuval worked on extending MBSL for handling dependency relations. Preliminary results indicate that it may be possible to use MBSL for producing a partial parse for dependencies, using a method which was originally aimed at phrasal structures. This issue is still under study.

He had interesting discussions with Sandra Kübler (research and teaching staff at University of Tübingen), who is working on her own memory-based parsing system. They exchanged

views and comments, and are likely to continue doing so. As part of the postgraduate course on statistical methods for parsing, Yuval presented MBSL to the group members.

Another line of his research is a view of statistical NLP systems as samples of statistical variables. As such, their performance is itself a statistical variable, and we can study its distribution and correlation properties. Yuval presented a preliminary results of his study in ACL-2001 workshop on Evaluation for Language and Dialogue Systems (Krymolowski 2001). The results indicate that the distribution can indeed provide information about similarity between corpora. As the width of the distribution is related to the sensitivity of the system to noise in the training data, this approach provides information about systems which goes beyond mere comparison of performance data.

Such study, which is more basic in nature, 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. We also need to estimate the probability of getting similar or worse/better performance on data from different genres. Along this line, we need at first to have a measure for data similarity related, in turn, to the task at hand. These issues are dealt with in the presented paper.

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

5.7.7 Invited researcher: Wouter Jansen from Groningen

Wouter Jansen is a PhD student in the departments of Linguistics and Humanities Computing at the University of Groningen. His supervisors are Dicky Gilbers and John Nerbonne. He visited Tübingen between Sept. 2001 and March 2002.

Wouter's main research interests are 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).

Wouter's stay in Tübingen offered him the opportunity to learn about finite-state implementations of Optimality Theory (Prince & Smolensky 1993) from Dale Gerdemann, which directly improved his work on optimality-theoretic models of English phrase and (nominal) compound stress. During his time in Tübingen, Wouter 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.

5.7.8 Dale Gerdemann, Site Coordinator

Dale Gerdemann has coordinated LCG research at Tübingen. 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.

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 resulted in a publication in the journal *Grammars* (van Noord & Gerdemann 2001).

5.7.9 Related Tübingen Researchers

Prof. Dr. Erhard Hinrichs and Sandra Kübler, M.A. are working on the application of symbolic machine learning techniques to parsing (Hinrichs, Kübler, Müller & Ule 2002, Kübler & Hinrichs 2001*a*, Kübler & Hinrichs 2001*b*, Kübler 2001). They are using a similarity-based approach to extend partial chunk parses into complete tree structures, including function-argument structure.

They also organized a workshop on "Machine Learning Approaches in Computational Linguistics" which has since been held at the ESSLLI 2002 summer school in Trento.

During the stay of Yuval Krymolowski, who has been working on extending MBSL for handling dependency relations, he and Sandra Kübler had fruitful discussions on different memory-based techniques and their application to (partial) parsing. This collaboration will be continued in the future.

Tylman Ule (staff member) participated in the seminar on statistical parsing methods along with Franck Thollard and Yuval Krymolowski. He has 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.

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 Master's 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, he has presented at the international Finite State Methods in Natural Language Processing workshop at ESSLI in Helsinki (Vaillette 2001).

References

- 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.
- Ashby, S., J. Carson-Berndsen & G. Joue (2001), A testbed for the development of multilingual phonotactic descriptions., in ‘Proceedings of Eurospeech 2001, Aalborg, September’.
- Bohan, A., E. Creedon, J. Carson-Berndsen & F. Cummins (2001), Application of a computational model of phonology to speech synthesis, in ‘Proceedings of AICS 2001, Maynooth, September’.
- 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.
- Carson-Berndsen, J. & D. Gibbon (2001), Visualising lexical prosodic representations for speech applications, in P.McKevitt, S.Nuallein & C.Mulvihill, eds, ‘Language, vision and music, Readings in Cognitive Science and Consciousness, Advances in Consciousness Research, AiCR, Amsterdam, The Netherlands’, John Benjamins Publishing Company, Philadelphia, USA.
- Carson-Berndsen, J., G. Joue & M. Walsh (2001), Phonotactic constraint ranking for speech recognition, in ‘Computational Linguistics in the Netherlands 1999. Selected Papers from the Eleventh CLIN Meeting’.
- Carson-Berndsen, J. & M. Walsh (2001), Defining constraints for multilinear speech processing, in ‘Proceedings of Eurospeech 2001, Aalborg, September’.
- Cummins, F. (2001a), ‘On synchronous speech’, *Acoustic Research Letters Online*, <http://ojsps.aip.org/ARLO/>.
- Cummins, F. (2001b), ‘Reducing expressive variation in speech with synchronous speech’, *Journal of the Acoustical Society of America* **109**(5(2)), 2416–2417.
- Cummins, F. & R. Deb. (2001), ‘Using synchronous speech to minimize variability in pause placement’, *Proceedings of the Institute of Acoustics* **23**(3), 201–206.
- 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.
- Déjean, Hervé (2002), ‘Learning linguistic rules and their exceptions’, *Journal of Machine Learning Research* **2**, 669–693.
- Eisner, Jason (2000), Bilexical grammars and their cubic-time parsing algorithms, in H.Bunt & A.Nijholt, eds, ‘Advances in Probabilistic and Other Parsing Technologies’, Kluwer Academic Press.

- Gaussier, Eric, Cyril Goutte, Kris Popat & Francine Chen (2002), A hierarchical model for clustering and categorising documents, *in* 'Proceedings of the European Colloquium on Information Retrieval (ECIR 2002)', Glasgow, UK.
- 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.
- Hammerton, James & Erik F. Tjong Kim Sang (2001), Combining a self-organising map with memory-based learning, *in* W.Daelemans & R.Zajac, eds, 'Proceedings of CoNLL-2001', Toulouse, France, pp. 9–14.
- Hammerton, James, Miles Osborne, Susan Armstrong & Walter Daelemans (2002), 'Introduction to special issue on machine learning approaches to shallow parsing', *Journal of Machine Learning Research* 2(March), 551–558.
- 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 Klassifikation', Classification, Automation, and New Media.
- 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.
- Joue, G. & J. Carson-Berndsen (2001), An embodiment paradigm for speech recognition systems, *in* 'Proceedings of Eurospeech 2001, Aalborg, September'.
- Koeling, Rob (2002), Dialogue-Based Disambiguation: Using Dialogue Status to Improve Speech Understanding, PhD thesis, Maths and Sciences, Rijksuniversiteit Groningen.
- Konstantopoulos, Stasinou 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/essli01.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), Tüsbl: A similarity-based chunk parser for robust syntactic processing, *in* 'Proceedings of HLT', San Diego, Cal.
- Miller, Stefan (1999), *Deutsche Syntax deklarativ: Head-Driven Phrase Structure Grammar für das Deutsche*, number 394 *in* 'Linguistische Arbeiten', Max Niemeyer Verlag, Tübingen.
- 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.

- Nenova, N., G. Joue, R. Reilly & J Carson-Berndsen (2001), Sound and function regularities in interjections, *in* ‘Proceedings of Disfluency in Spontaneous Speech, Edinburgh, Scotland, August’.
- Nenova, N. & R. G. Reilly (2001), A taxonomy of discourse particles in spoken language, *in* ‘Proceedings of 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.
- Prince, Alan & Paul Smolensky (1993), Optimality Theory: Constraint interaction in generative grammar, Technical Report 2, Center for Cognitive Science, Rutgers University.
- 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’.
- Srinivasan, Ashwin (2001), *Aleph*, <http://www.comlab.ox.ac.uk/oucl/research/areas/machlearn/Aleph/>.
- Thollard, Franck (2001*a*), 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 (2001*b*), 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. (2001*a*), Memory-based clause identification, *in* W.Daelemans & R.Zajac, eds, ‘Proceedings of CoNLL-2001’, Toulouse, France, pp. 67–69.
- Tjong Kim Sang, Erik F. (2001*b*), Transforming a chunker to a parser, *in* ‘Computational Linguistics in the Netherlands 2000’, Tilburg, The Netherlands.
- Tjong Kim Sang, Erik F. (n.d.), ‘Memory-based shallow parsing’. (submitted to the Journal of Machine Learning Research).
- Tjong Kim Sang, Erik F. & Hervé Déjean (2001), Introduction to the conll-2001 shared task: Clause identification, *in* W.Daelemans & R.Zajac, eds, ‘Proceedings of CoNLL-2001’, Toulouse, France, pp. 53–57.
- Vaillette, Nathan (2001), Logical specification of transducers for nlp, *in* ‘Proceedings on FSMNLP 2001’, University of Helsinki.
- van Noord, Gertjan & Dale Gerdemann (2001), ‘Finite state transducers with predicates and identity’, *Grammars* 4(3), 263–286.