

Learning Computational Grammars

TMR Project Nr. ERBFMRXCT980237

2nd Annual Report

John Nerbonne and Erik Tjong Kim Sang

December 8, 2000

Summary

This document describes the second year's progress of the TMR Project *Learning Computational Grammars* (LCG). In brief, LCG now has a full complement of postdocs, and work is ongoing 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 are targeting their various learning technologies on the task of learning noun phrases in free text. The industrial partner, Xerox, is exploring an application, and one of the academic sites has focused on a comparison to linguistic and psycholinguistic accounts of learning (Geneva). LCG created a task description and an attendant training and testing set, which was the focus of a public meeting, held in conjunction with the Special Interest Group in Natural Language Learning and the European Chapter of the Association for Computational Linguistics in Bergen in June 1999. A mid-term meeting was held in Dublin in November 1999.

A highlight of the second year has been that the network is now responsible for three of the world's four best results in the on the recognition of simple noun phrases in text (cf. reports of Koeling, Tjong Kim Sang and Déjean at the Dublin meeting).

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:

Antwerp	Instance-based Learning
Dublin	Neural Networks
Groningen	MDL, Random Fields (Maximum Entropy), ILP, Neural Networks
SRI	Genetic Algorithms, Maximum Entropy, ILP
Tuebingen	Theory Refinement
Xerox	EBL

The ISSCO group at Geneva concentrates on complementary work on the linguistic and psycholinguistic subtleties of noun phrases.

2 Group Meetings

We had two network meetings in the second year of the network. The first was collocated with the Conference of the European Chapter of the Association for Computational Linguistics (EACL'99) and the workshop on Computational Natural Language Learning (CoNLL-99) in Bergen, Norway in June 1999. The second network meeting was the mid-term meeting held in Dublin, Ireland in November 1999.

2.1 CoNLL Meeting, Bergen, June 1999

The Third Workshop on Computational Natural Language Learning (CoNLL-99, <http://lcg-www.uia.ac.be/conll99/>) was organised by two of the network postdocs (Osborne and Tjong Kim Sang). Because of the delay in appointing people at the different sites, this was the first time that every site was represented by a postdoc. The final session of CoNLL-99 was reserved for presentations about a shared task, recognising arbitrary noun phrases. This is the first shared task of the network but participation in this task was open to non-network participants as well. Unfortunately, the two network organisers were the only ones able to present results for the shared task at the time of the workshop. We have used the remainder of the session for giving the other network participants the opportunity to introduce themselves and their work.

- 15.30 – 16.00 MDL-based DCG Induction for NP Identification
Miles Osborne (Groningen)
- 16.00 – 17.00 NP Identification Session
James Hammerton (Dublin)
Rob Koeling (Cambridge)
Nicola Cancedda (Grenoble)
Hervé Déjean (Tübingen)
Erik Tjong Kim Sang (Antwerp)

After this session, we had a brief internal meeting. The main discussion points were the next meeting in Dublin, opportunities for visits to other sites and the plans for publishing a book with the work of the network participants.

2.2 Mid-Term Meeting, Dublin, November 1999

This meeting was spread over two days. At the first day, all persons employed on the network (seven postdocs and two PhD students) gave a presentation of their work. Additionally there was a presentation of network coordinator John Nerbonne, another by Tjong Kim Sang and Osborne regarding the definition of the shared tasks of the network and an invited talk by Marshall Mayberry of the University of Texas. At the end of the day, we discussed future meetings, a possible extension of the network to account for the late start and an opportunity to send the postdocs and PhD students to XRCE, Grenoble for a course on finite state tools.

The second day was a more formal mid-term review of the project. This review was chaired by Christiane Bernard, DG XII. At this review there were talks by the meeting chair, the local network coordinators and all persons employed on the network.

We include the programme of the two days here. Overhead sheets of most of the talks can be obtained via <http://lcg-www.uia.ac.be/lcg/meetings/9911.html> .

Monday 15 November 1999

- 09:00 Coffee and Welcome – Ronan Reilly, Dublin
- 09:15 LCG Network: Motivation, Initiation, Progress
John Nerbonne, Groningen
- 09:45 Shared Task Definition
Erik Tjong Kim Sang, Antwerp
Miles Osborne, Groningen
- 10:15 Investigation of SARDSRN, a Connectionist Architecture
for NLP, in Noun-Phrase Identification
James Hammerton, Dublin
- 10:45 Break
- 11:00 MDL-based Learning
Miles Osborne, Groningen
- 11:30 LCG at SRI Cambridge, UK
Rob Koeling, SRI Cambridge
- 12:00 Applying Explanation-Based Learning to
Lexical Functional Grammars
Nicola Cancedda, Xerox, Grenoble
- 12:30 Lunch
- 13:30 Memory-Based Learning
Erik Tjong Kim Sang, Antwerp
- 14:00 LCG in Tübingen, Germany
Hervé Déjean, Tübingen
- 14:30 Genetic Algorithms
Anja Belz, SRI Cambridge
- 15:00 Break

- 15:30 Guest Presentation
Marshall Mayberry, University of Texas
Research on Language Learning in Neural Nets
- 16:30 Error Analysis
Adelina Hild, Geneva
- 16:45 Inductive Logic Programming
Stasinos Konstantopoulos, Groningen
- 17:00 General Discussion

Tuesday 15 November 1999: Mid-Term Project Review

- 09:00 Informal Welcome by Ronan Reilly, Dublin
- 09:15 Opening
Christiane Bernard, DG XII (Meeting Chair)
- 09:45 Coordinator's Report
John Nerbonne, Groningen
- 10:45 Break
- 11:15 Discussion of Coordinator's Report
- 11:30 Tour de Table (scientists-in-charge)
John Nerbonne, Groningen
Dale Gerdemann and/or Erhard Hinrichs, Tübingen
Walter Daelemans, Antwerp
David Milward, SRI Cambridge
Ronan Reilly, Dublin
Christer Samuelsson, Xerox
Susan Armstrong, Suissetra
- 12:30 Lunch
- 14:00 Young Researchers' Reports
Postdocs
James Hammerton, Dublin
Miles Osborne, Groningen
Nicola Cancedda, Xerox, Grenoble
Erik Tjong Kim Sang, Antwerp
Hervé Déjean, Tübingen
Anja Belz, SRI Cambridge
Adelina Hild, Geneva
PhD students
Rob Koeling, SRI Cambridge
Stasinos Konstantopoulos, Groningen
- 15:30 Break
- 16:00 General Discussion
- 17:00 Closing

3 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, 1999 – March 31, 2000.

3.1 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.

3.1.1 Erik Tjong Kim Sang, Postdoc

In the second project year, Erik has continued with work on base noun phrase identification and started working on recognising arbitrary noun phrases and finding arbitrary base phrases. He has used this year to explore the usability of system combination: a technique which attempts to boost performance by combining the results of different learning algorithms to the same problem. Erik has both applied this to the mentioned tasks by applying memory-based learning to different variants of the same task and by using different learning techniques.

For base noun phrases, Erik has completed a study on the influence of the output representation of phrases on recognition performance (joint work with Jorn Veenstra, published in EACL'99). As a followup, he examined how results obtained with these representations can be combined (paper accepted for NAACL 2000). Finally, he investigated the possibility of combining the results of different machine learning algorithms to this problem (joint work with Walter Daelemans, Hervé Déjean, Rob Koeling, Yuval Krymolowski, Vasin Punyakanok and Dan Roth, paper accepted for COLING 2000).

For arbitrary noun phrase recognition, Erik designed a method for applying memory-based learning to this problem (talk presented at CoNLL-99). He applied a data-representation-based system combination to this problem as well (results published in the accepted NAACL 2000 paper). For arbitrary base phrases, he has examined different methods of applying memory-based learning to this problem (paper submitted to ACL2000).

Apart from the noun phrase combination joint work with the sites in Tübingen (Déjean) and Edinburg (Koeling), Erik has worked together with Adelina Hild from Geneva on base noun recognition error analysis. Outside of the project Erik has done some combination work for phonemic data analysis (joint work with Véronique Hoste, Walter Daelemans and Steven Gillis, paper accepted for ICML-2000). Erik has co-edited the proceedings of the CoNLL-99 (joint work with Miles Osborne). Currently he is involved in the organisation of CoNLL-2000 as the shared task coordinator.

3.1.2 Walter Daelemans, Coordinator

During the report year, a new project was started on machine learning of language funded by CELE, the research centre of S.A.I.L. trust (Ieper). In this project, a postdoc (Jakub Zavrel) and a PhD student (Anne Kool) study the integration of evolutionary computing and neural networks in statistical and memory-based learning of language, and issues of modularity and combination in machine learning systems. Several projects in the machine learning of language mentioned in the first annual report were continued by Véronique Hoste (Linguaduct), Guy De Pauw, Masja Kempen, and Helena Taelman. CNTS attracted three additional PhD students (with their own funding) working on related topics (statistical and machine learning approaches to NLP).

CNTS (co-directed by Daelemans and Gillis) became a member of MLNET, the European network of excellence on machine learning, and coordinates a five year extension of CLIF (Computational Linguistics in Flanders), a government-sponsored research community 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 this and related projects, especially Sabine Buchholz and Jorn Veenstra.

Relevant publications of Daelemans and other CNTS participants can be found in the bibliography.

3.1.3 Training activities

Erik has taken part in the tutorial *Natural Language Learning with the Maximum Entropy Framework* by Adwait Ratnaparkhi at the EACL'99. He has developed and taught a Perl course for our industrial partner S.A.I.L. (L&H). Apart from that he has presented four undergraduate lectures in Antwerp and two lectures in Ieper (S.A.I.L.). The lectures covered statistical and connectionist natural language processing. Currently Erik is supervising one student who is writing a Masters thesis on connectionist language processing.

One day in the week, Erik works in the Induction of Linguistic Knowledge group in Tilburg, The Netherlands. In his case, the cooperation with Tilburg has resulted in one joint conference paper (EACL'99 with Jorn Veenstra).

The Antwerp site was visited by Sandra Kübler from Tübingen in December 1999 for a tutorial in memory-based learning. The tutorial was presented by Walter Daelemans.

3.2 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.

3.2.1 Summary of LCG Project Activities at SRI

LCG Project activities at SRI International during the reporting period included research on shared project tasks and related grammar learning problems, submission of papers to international conferences and workshops, regular LCG project meetings, in the form of seminars, a reading group and invited talks, creation and maintenance of extensive project-related web pages, and attending project-related conferences, workshops, seminars and meetings.

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

The SRI LCG group now have extensive LCG project web pages, providing information about LCG research at SRI, and links to other web pages containing information about natural language learning methods and NLL and NLP tools (<http://www.cam.sri.com/tmr>).

During the reporting period, SRI LCG project researchers attended conferences and meetings, including EACL'99 in Bergen, and the LCG project meetings in Bergen (June) and Dublin (November). They also presented talks locally and at project meetings 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 an important 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 include a seminar series on computational grammar formalisms, submission of papers to refereed conferences and journals, and visits to other LCG project partners.

3.2.2 Anja Belz, Postdoc

Anja's research involves the following activities:

- *Genetic Algorithms for Finite-State Grammar Learning*: As the first research subject on the LCG Project, Anja adapted a previously developed genetic algorithm for learning generalised finite-state grammars for NP-chunking. The training and testing data was generated using existing grammars in SRI's Highlight system. The overall result was that meaningful generalisation can be achieved, but that the cost of using a GA to construct and generalise grammars is too high for wide-coverage grammar development.
- *Treebank Grammars*: The term "treebank grammar" has come to mean PCFGs that are extracted directly from bracketed and annotated corpora. The number of different bracketings in corpora tends to be very large, and therefore extracted grammars tend to have large numbers of rules (e.g. 33,000 rules from WSJ Corpus, Sections 02-21). Anja investigated the performance of such treebank PCFGs on the task of parsing unseen test sentences, and the effect different methods of grammar reduction have on parsing performance. Of particular interest were the fact that different types of rules are affected differently by reduction methods (e.g. parsing performance is better on lower-level constituents).
- *Addition of Structural Context to PCFGs*: As a follow-on project from the treebank grammar research, Anja is currently investigating different sources of low-cost structural information directly derivable from treebanks, such as the identity of parent labels, grammatical role and the depth of embedding of rule applications, and how to incorporate such context into treebank PCFGs.
- *Shared Task 'Arbitrary Chunking'*: In a subproject of the previous, Anja is adapting a structure-sensitive PCFG for the task of arbitrary chunking (as defined by Tjong Kim Sang for CoNLL-2000).

Anja presented the following talks: *Genetic Search Algorithms for NP Learning* (16 July 1999), TMR-LCG Project Presentation, SRI International, Cambridge, *Learning Finite-State Noun Phrase Grammars* (4 August 1999), Local LCG Seminars, SRI International, Cambridge, and *NP Learning with Treebank Grammars and Genetic Algorithms* (16 November 1999), TMR-LCG Project Meeting, University College Dublin.

Apart from her research activities, Anja administrated and organised the local LCG Seminar Series, designed and maintained local TMR-LCG web pages, attended EACL'99, CoNLL-99, and Royal Society Meeting on Computers, Language and Speech.

In the next project year the research of Anja will focus on work on the shared task for the CoNLL Workshop at ICGI-2000 and the continuing investigation of incorporating structural context into PCFGs. She already has five scheduled presentations: a project talk at the TMR-LCG Project Meeting at XRCE, Grenoble (*Adding Structural Context to PCFGs*, 19 May 2000), invited talks at the Department of Computer Science, UCD (12 May 2000), ITRI, Brighton University (8 June 2000), and the Computer Lab, Cambridge University (15 June 2000), and a refereed conference talk at the SIGPHON Workshop at COLING 2000 (6 August 2000).

3.2.3 Rob Koeling, PhD student

Rob's research involved the following activities:

- As a first exercise, Rob performed experiments on base NP chunking using MaxEnt models. Building on results from baseNP chunking experiments, Rob started experiments for shared task 1): Annotating sentences with parentheses marking NP boundaries. First results were presented at Dublin meeting (November 1999). A project report describing method and results for both series of experiments is forthcoming.
- Some research was done on feature selection and smoothing techniques for MaxEnt models.

His publications include *Using Maximum Entropy Modelling for Contextual Interpretation of Answers*, TST Technical Report 99 (September 1999), *Using dialogue information for parsing in a spoken dialogue system*, to appear in Proceedings of Gotalog, Gothenburg, Sweden, 2000, and *Applying System Combination to Base Noun Phrase Identification*, to appear in Proceedings of COLING 2000, Saarbrücken, Germany, 2000 (joint work with Tjong Kim Sang, Daelemans, Déjean, Krymolowski, Punyakanok and Roth).

Rob gave a local presentation and introduced a paper at weekly LCG seminars. He also presented *MaxEnt NP chunking* at the TMR-LCG Project Meeting, University College Dublin (16 November 1999), and *A Maximum Entropy Model for adding context in a spoken dialogue system*, Computational Linguistics in the Netherlands. Clin X, University of Utrecht (10 December 1999).

In this second project year, Rob visited LCG partners Groningen University and Tübingen University. He attended EACL'99 and CoNLL-99 in Bergen, Norway.

Rob's involvement with TMR-LCG will end 1 August 2000. For the remaining months he has planned a project talk at the TMR-LCG Project Meeting at XRCE, Grenoble (19 May 2000) and a poster presentation at Gotalog, Gothenburg, Sweden (15-17 May 2000). Furthermore, he will contribute to the shared task of arbitrary chunking as defined for the CoNLL-2000 workshop and give a MaxEnt tutorial at the local Statistical NLP seminar series.

3.2.4 David Milward, Coordinator

David Milward attended the Dublin LCG meeting and participated in the local LCG paper reading and seminar series.

3.2.5 Related Cambridge researchers

Dr Stephen Pulman (Principal Scientist at SRI and Reader at University of Cambridge Computer Laboratory) continues to take an interest in LCG activities, in particular those concerning inductive logic programming. He invited Dr James Cussens to give a seminar on ILP at SRI last year.

Dr Ted Briscoe (Lecturer, University of Cambridge Computer Laboratory) continues to be actively involved in the LCG Project. He attends LCG seminars, reading groups and invited talks, and has recently completed the project on improving the GR annotation of a test corpus (from the SUSANNE Corpus) to cover NP internal structure which will be used for evaluation purposes by project members (<http://www.cogs.susx.ac.uk/lab/nlp/carroll/greval.html>).

Sylvia Knight (part-time researcher at SRI and Doctoral Researcher and Tutor at University of Cambridge Computer Laboratory) also continues to take an active interest in the LCG Project. During the reporting period she attended group meetings and introduced two papers to the LCG reading group.

Other members of SRI's NLP research group in particular Dr Ian Lewin and Dr Richard Tucker, have been taking an active interest in the LCG project, participating in the local LCG seminars and reading group. SRI had two project-related invited speakers during this year: James Cussens, *Introduction to Inductive Logic Programming* (April 2000) and Miles Osborne, *MDL-based Learning* (November 1999).

3.2.6 Training activities

Rob Koeling attended the tutorial *Natural Language Learning with the Maximum Entropy Framework* by Adwait Ratnaparkhi at EACL'99 and he gave a tutorial on Maximum Entropy modelling at fellow LCG partner Tübingen. He has visited LCG partner Groningen University and he will attend the XFST course at XRCE in Grenoble in May 2000.

3.3 Dublin

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

3.3.1 James Hammerton, Postdoc

At the time of the last report, James was investigating the use of the SARDSRN architecture [Mayberry and Miikkulainen, 1998] for the NP task and the use of the (S)RAAM

[Callan and Palmer-Brown, 1997] for representing the output from the SARDSRN. The (S)RAAM has since been abandoned as it was found to scale poorly with the size of the training set.

To devise a more scalable representation, the SARDNET [James and Miikkulainen, 1995] has been investigated. The SARDNET has a high theoretical capacity. Given N units, N sequences of length 1 can be represented plus $N \times (N - 1)$ sequences of length 2 plus ... plus $N!$ sequences of length N can be represented. Investigations confirmed that this capacity can be exploited. A set of sequences were compiled from sections 15 to 18 of the WSJ corpus comprising of the part of speech tags for all sentences < 20 words long, with bracketing to indicate noun phrases. This yielded 3115 distinct sequences. A 100 unit SARDNET was trained successfully to map these sequences. When trained on 1565 of the sequences it produced distinct patterns for the entire set. It thus appears that the SARDNET can form dense maps of the sequences and generalises very well. A simple recurrent network was successfully trained to decode SARDNET representations on a set of 125 active/passive sentence pairs, but there has not yet been any success in decoding larger data sets. To test whether SARDNET representations can be used as outputs in connectionist systems, a feedforward network (FFN) was trained to take the SARDNET representation of active sentences and output the equivalent passive sentences. Generalisation was poor however. When trained on 65 of the 125 possible transformations it only generalised to 6 of the rest. Unfortunately noise disrupts the information encoding the order of symbols in the sequence. Some method of making the SARDNET representations more robust is needed to solve this. This line of work has been suspended to concentrate on the LCG tasks, however.

Regarding the SARDSRN, the best result is that a SARDSRN with 2 49-unit SARDNETs with 196 hidden and context units, was trained on 300 sequences of length < 10 from the WSJ corpus for 1000 iterations, over a period of 16.5 hours. The SARDSRN was outputting a representation of the bracketed sentence (with place markers for the words). 296 sentences were processed perfectly. However only 2 out of 5 runs approached this performance. Using the > 3000 sentences of max length 20 from sections 15 to 18 of the WSJ corpus, would take at least 2 weeks, by extrapolation. Thus the training is too long as well as being unreliable. Alternative approaches to the problem are now being considered, such as splitting into into a set of smaller tasks for an ensemble of networks, and considering new training algorithms.

Finally, James attended the Computational Natural Language Learning Workshop (CoNLL-99) at the University of Bergen, Bergen, Norway on the 12th June. This doubled as a LCG meeting, and James gave a short talk about his LCG work. A talk about using SARDNET as a representation was presented at the EmerNet: International Workshop on Emergent Neural Computational Architectures Based on Neuroscience, University of Edinburgh, UK (11 December 1999, <http://www.his.sunderland.ac.uk/worksh2/>). An abstract and talk on the same topic was presented at the Emergent Computing: Self Organising Systems – Future Prospects for Computing workshop at the Manchester Conference Centre, UMIST, Manchester, UK (28-29 October 1999, <http://images.ee.umist.ac.uk/emergent/>). Furthermore, James gave a talk about his PhD at Cognitive Science for the New Millennium (see conference details below) and presented a paper entitled *Holistic Symbol Processing*, [Hammerton, 1999], to the AICS'99 Conference at University College Cork.

3.3.2 Ronan Reilly, Coordinator

During the last year of the project, Ronan was involved in a number of activities both NLP-related and related to cognitive science. Ronan was involved in the organisation of the conference Cognitive Science for the New Millennium (16-17 May 1999). This conference was held to mark the launch of the new MA/MSc degree in Cognitive Science at UCD. It was attended by more than forty cognitive science researchers from Europe and the US. The highlights of the conference were the invited talks from Professor James McClelland of Carnegie-Mellon University and Professor Paul Smolensky of Johns Hopkins University. Both speakers are acknowledged pioneers in the field of cognitive science. The conference ended with a round-table discussion on the prospects for cognitive science in the next millennium, with Prof. James McClelland, Prof. Paul Smolensky, Prof. Ciaran Regan, Dr. Maria Baghramian, chaired by Dr. Ronan Reilly.

In September 1999, he taught a course on connectionist grammar learning at Tübingen for the Tübingen-Sofia Graduate Programme in Computational Linguistics and Represented Knowledge (CLaRK) Summer School. CLaRK provided a joint teaching/research facility where graduate students primarily from Bulgaria and CEE pursued their research in computational linguistics and knowledge representation. The programme drew upon the complementary strengths of its host institutions, the Seminar fuer Sprachwissenschaft of the Eberhard-Karls-Universitaet, Tübingen, Germany and the Linguistic Modelling Laboratory of the Bulgarian Academy of Sciences, Sofia, Bulgaria. The scientific coordinators of the programme were Erhard Hinrichs (SfS), Paul John King (SfS) and Kiril Ivanov Simov (LML).

He presented a paper on the issue of systematicity in grammar learning at the Pacific Rim NLP conference in Beijing (October 1999). He gave an invited talk at the Cognitive Science Colloquium, SUNY Buffalo, on the topic of language evolution and development (December 1999). Ronan has also had 2 publications during this period [Reilly, 1999, Reilly, 200x].

3.3.3 Related Dublin researchers

Dr. Arthur Cater has been involved in the organisation of an ongoing series of computational linguistics seminars at UCD. Speakers have been invited from a wide range of research groups, both in Europe and the US. Among the invited speakers was Anja Belz, the LCG postdoc based at SRI Cambridge.

Dr. Julie Berndsen gave an invited talk to the Royal Society in London on her work on finite state models of phonology. Dr. Berndsen has also been in contact with Anja Belz with a view to possible future collaboration.

3.3.4 Training activities

James was involved in the following training activities:

- He proposed and co-supervised (with Ronan Reilly) a summer project where the student, Paul Heffernan, investigated whether some problems with a representation known as the Sequential RAAM [Hammerton, 1998] could be alleviated. He also proposed and co-supervised (again with Ronan) Paul's final year project, which investigated whether a representation known as the Bi-coding RAAM (BRAAM) [Adamson and Damper, 1999] can represent logical terms, and whether a FFN can be trained to take 2 BRAAM representations of logical terms and unify them. For a large set of simple terms, the BRAAM and FFN successfully learned their tasks.
- He attended some of the more psychologically oriented courses on the Cognitive Science MSc course at UCD.
- He arranged a research visit for 28 May – 9 July 2000 at the University of Antwerp. One line of work being pursued for this is investigating whether a neural network can perform the classifications in memory-based learning (MBL). Pilot studies are promising with a FFN achieving results comparable to MBL on base noun-phrase chunking using the current POS tag and previous POS tag as features. It is hoped to scale this up to using more features, and to doing the full NP bracketing task.

3.4 ISSCO, Geneva

This site employs one postdoc. This subsection reports on her research and training activities and gives an overview of the work of other people working at this site.

3.4.1 Adelina Hild, Postdoc

Adelina joined ISSCO and began her work on the project in June, 1999. For the last year, ISSCO's contribution to the project has had two focal points. Firstly, work focusing on the acquisition and processing of NPs by humans and related investigation of recursive NPs with emphasis on nominal phrases headed by deverbal nominals. Secondly, the site is preparing POS-tagged multilingual corpora (English, German and French), which is to be made available to the other participants in the project for further experimental work.

LCG project-related activities:

- Review of available experimental research on the acquisition and processing of phrase structures by humans, with specific emphasis on NP acquisition;

- Error analysis of the results from the NP chunking experiment carried out by Erik Tjong Kim Sang at the Antwerp site [Tjong Kim Sang and Veenstra, 1999]. The aim of the analysis is to develop an annotation standard for automatic parsing for the grammatical structures that pose processing problems. The proposed classification of errors was developed with reference to other available error analysis schemes, most notably TSNLP. The TSNLP scheme suggested various concepts around test suite design, which were deemed appropriate for a later stage of the project. In connection with this Adelina met two of the participants in the TSNLP project. Further work on automatic error detection has to be carried out together with the Antwerp site and a planned visit to the site is forthcoming;
- Tagging the multilingual 'EU Parliamentary Debates' corpus and developing the German and English lexicons and morphology. This task was carried out in conjunction with other ISSCO members – Sabine Lehmann and Pierrette Bouillon – whose role has been instrumental in developing the linguistic modules for the three languages chosen for the corpus. To date, the English and German morphology have been updated and refined (the French morphology was in a ready-to-use state), which allowed to undertake the tagging of the respective parts of the corpus. The next immediate step in this direction is evaluating and fine tuning the morphology for the three language modules. During a visit to the University of Tübingen, Adelina met with the team working on the VERBMOBIL project to assess the possibility of syntactically annotating the corpus, which might be required for the second phase of the LCG project.
- She attended the workshop on Computational Natural Language Learning in the course of the EACL'99 conference, Bergen, Norway;
- Adelina gave a talk on error-analysis results from the above referenced machine-learning experiment at the Dublin meeting of the project participants, November, 1999.

3.4.2 Related Geneva researchers

Pierrette Bouillon and Sabine Lehmann have been involved in the preparation of the corpus. Pierrette Bouillon works on another learning project: Action de Recherche Partagée, réseau Francil, "Acquisition automatique d'éléments du Lexique Génératif pour améliorer les performances de systèmes de recherche d'informations". The aim of the project is to propose an Inductive Logic Programming learning method with aim at automatically extracting semantically related Noun-Verb pairs from a corpus in order to build up semantic lexicons based on Generative Lexicon principles. Sabine Lehmann will attend the workshop on Finite State Tools, organised by the Xerox site in Grenoble from the 15th to 18th of May 2000. It will present an introduction to to some of the Xerox's Finite State Tools, in particular xfst and lexc. These tools have been installed at ISSCO and will be used for processing German compounds.

3.4.3 Training activities

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

Adelina visited the Tübingen site in December, 1999, where she looked in detail at ALLiS (the learning system used there) and the local results from the first learning task: identifying base NPs.

3.5 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.

3.5.1 Nicola Cancedda, Postdoc

TMR-LCG research at Xerox focuses on applying Explanation-Based Learning for specialising LFG grammars. The activities conducted in the period under consideration include:

- **Treebank annotation.** Design and implementation of the software for annotating LFG treebanks with the additional information required to perform EBL on them.
- **Benchmarking.** Design and implementation of the software for assessing the performance of specialised LFG grammars through ten-folded cross-validation experiments. Performance indices include coverage, ambiguity reduction and speedup, both as global averages and distributed according to sentence length.
- **Preliminary experiments.** Preliminary experiments with a very basic form of EBL, consisting in retaining all and only the portions of the grammar used to parse the training corpus ([Cancedda and Samuelsson, 2000]).
- **Benchmark validation.** Validation of the performance assessment procedure through application to the grammars learned by the preliminary experiments.
- **Full-coverage architecture.** Simulated experiments with an extended, two-stage architecture.
- **EBL experiments.** Design and implementation of the full EBL experiments. The problem is formulated as the search for the specialised grammar which optimises a

function combining ambiguity reduction, coverage and grammar size. A representation of the corpus suitable for this formulation has been designed.

- **Assessment of EBL results.** Assessment of the performance of the grammars specialised by means of full EBL through the benchmarking procedure established in 3.5.1.

3.5.2 Christer Samuelsson, Coordinator

In addition to supervising the work carried out in the TMR/LCG project by Nicola Cancedda, Christer Samuelsson has conducted theoretical research on stochastic grammars and applied this to statistical dependency parsing, resulting in two conference publications ([Samuelsson, 2000a, Samuelsson, 2000b]). In addition to this, he has lead a rather applied research effort where machine-learning and statistical techniques were used to improve optical character recognition.

3.5.3 Related Grenoble researchers

Andreas Eisele worked on a stochastic model of syntactic dependencies for the resolution of ambiguity in LFG analyses. Each potential analysis in a f-structure chart is scored on the basis of its Predicate values and their interrelations. The parameters of the stochastic model are trained from a “shallow analysis” of a large corpus from a general domain, plus a small number of manually disambiguated LFG-analyses from the test domain. The work is described in [Eisele, 1999]. Additionally, he contributed to David Hull’s work on the acquisition of bilingual lexical knowledge from translation examples.

Eric Gaussier designed an unsupervised method to learn suffixation operations of a language from an inflectional lexicon. This method, based on hierarchical agglomerative clustering techniques and maximum likelihood optimisation, also leads to the development of a stemming procedure for the language under consideration [Gaussier, 1999].

An important research problem in machine-assisted translation is the automatic extraction of terminology units and their translations from existing translated texts. At XRCE, David Hull and others are developing hybrid systems which consist of a mixture of traditional rules-based NLP with statistical and machine learning algorithms. Terminology units are recognised in each language independently by defining regular expressions over sequences of part of speech tags. Words are aligned between the source and target language using a probabilistic alignment model which is fit via the EM algorithm. This alignment is incomplete because most source-target word pairs do not occur with sufficient frequency to align them correctly based solely on occurrence counts. A number of machine learning techniques, such as example-based learning, are currently being explored to produce accurate alignments of multi-word terms based on a partial alignment of their component words.

André Kempe worked on an entropy-based approach to segment a corpus into words, when no additional information about the corpus or the language, and no other resources such

as a lexicon or grammar are available. To segment the corpus, the algorithm searches for separators, without knowing a priori by which symbols they are constituted. Good results can be obtained with corpora containing “clearly perceptible” separators such as blank or new-line. The work was presented at the CoNLL-99 workshop in Bergen [Kempe, 1999].

3.5.4 Training activities

Nicola Cancedda attended the ACAI'99 summer school on Machine Learning in Chania, Greece, during July. He also attended the tutorial on *Natural Language Learning with the Maximum Entropy Framework* given by Adwait Ratnaparkhi at the EACL'99 in Bergen, as well as a tutorial on statistical techniques for NLP given by Christer Samuelsson for XRCE researchers.

Nicola Cancedda also co-animated a reading group on Machine Learning with about 20 participants.

3.6 Groningen

This site employs a postdoc and 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.

3.6.1 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 with an agreement to continue two months (in the summer of 2000) on the LCG project.

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

- Adaptation of the Sparkle DCG learner for NP identification. This task was straightforward, given the fact that the learner acquired NP rules already. The main change was allowing it to be trained on parsed corpora annotated with NP information.

- Induction of DCGs modelled as random fields. 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 submitted to 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 will be presented at CoNLL-2000 (Lisbon, September 2000) along with Tony Mullen in T. Mullen and M. Osborne, *Overfitting Reduction through Feature Merging for Maximum Entropy-based Parse Selection*. Tony Mullen also submitted a paper at the student ACL session in Hong Kong, 2000.
- For CoNLL-2000 Miles Osborne submitted a paper on *Shallow Parsing as Part-of-Speech Tagging*.
- 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 ACL2001.
- Also in conjunction with Rob Malouf, Miles Osborne has worked on the task of making maximum entropy efficient. Rob Malouf will submit a paper on the subject at CLIN 2000.

3.6.2 Stasinou Konstantopoulos, PhD student

Stasinou 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 a DCG for noun phrases in English from the Wall Street Journal annotated corpus (part of the University of Pennsylvania Treebank II project).

He is currently experimenting with non-recursive phrase chunking by syntactic tagging, which fits well with Aleph's single-predicate learning algorithm. A preliminary result has

been presented in the 10th Computational Linguistic in the Netherlands meeting. (Utrecht, 10 December 1999, <http://www.let.rug.nl/~vannoord/clin/>) [Konstantopoulos, 2000]. The goal is to (a) estimate the capabilities of the system when compared with other chunking-as-syntactic-tagging machine learning experiments, and (b) build experience on the system and the way that prior linguistic knowledge can be represented in the context of syntactic tagging, before attempting the more complex task of Full- (as opposed to Base-) NP parsing. The learning task is to be restricted to the domain of noun phrase syntax of Germanic languages, although care must be taken to allow extendibility to a full grammar.

Non-LCG academic activities include being one of the organisers of TABU-Dag 2000. TABU-Dag is an annual one-day conference on general linguistics, organised by the University of Groningen. This year's edition will take place on 16 June 2000, more at <http://www.let.rug.nl/tabu/>.

3.6.3 John Nerbonne, Coordinator

John Nerbonne continued investigations into phonological learning [Tjong Kim Sang and Nerbonne, 1999, Stoianov et al., 1999, Kleiweg and Nerbonne, 2000, Stoianov and Nerbonne, 2000] as well applications in computer-assisted language learning [Nerbonne and Dokter, 1999]. A third area of related activity was the application of unsupervised learning to the problem of dialect classification [Nerbonne et al., 1999b, Nerbonne et al., 1999a].

3.6.4 Related Groningen researchers

Tony Mullen 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 submitted to CoNLL-2000 (Lisbon, September 2000) along with Tony Mullen in T. Mullen and M. Osborne, *Overfitting Reduction through Feature Merging for Maximum Entropy-based Parse Selection*

Rob Malouf joined the Groningen group in July, 1999 as part of a university project focused on computational modeling of behavior, a collaboration between Computer Science, Computational Linguistics, Biophysics and Philosophy. He has focused on applying machine learning to a part of the LCG grammar task, namely word order in adjectival phrases [Malouf, 2000]. In addition he has work on efficient processing techniques [Malouf et al., 2000].

Gosse Bouma is a permanent staf member in Groningen who has recently attended LCG meetings and who has explored a task related to LCG's via machine learning, i.e. the grapheme-to-phoneme conversion problem [Bouma, 2000].

3.6.5 Training activities

Miles Osborne supervises Tony Mullen (PhD student), working on random field-based parse selection. He has prepared and delivered a course on Statistical Natural Language Processing here in Groningen. This is available at <http://www.let.rug.nl/~osborne/>

Stasinos Konstantopoulos attended the 11th European Summer School in Logic, Language and Information. (Utrecht, 9-20 August 1999, <http://ess11i.let.uu.nl>). He has also attended various talks on machine learning (e.g., by Ray Mooney, University of Houston, USA in Tilburg) and subjects of general interest (e.g. the weekly linguistics Colloquium in Groningen). He is also assisting Gosse Bouma by giving tutorials for a course in Natural Language Processing. (March - June 2000, see <http://www.let.rug.nl/gosse/nlp1/> for more details)

During this reporting period Nerbonne conducted a graduate level course on machine learning for Groningen graduate students in computational linguistics (including LCG member Konstantopoulos).

3.7 Tübingen

This site employs one postdoc. Here you will find an overview of his research and training activities as well as short reports on the work of the local coordinator as well as that of other related researchers.

3.7.1 Hervé Déjean, Postdoc

Before starting this position at the Seminar fuer Sprachwissenschaft, Hervé Déjean studied three years at Groupe de Recherche en Informatique, Image et Instrumentation de Caen (University of Caen). He defended his thesis untitled *Concepts et algorithmes pour la decouverte des structures des langues* in December 1998. The goal of this work was the extraction of syntactic structures from raw texts. During this period, he wrote four publications (three national and one international). He helped to the organisation of two conferences (CAPS'96 and CAPS' 98). During these three years, he also had a part-time position of teacher (90 hours per year) at the Computer Science Department (teaching Computer Science) and at the Linguistics Department (teaching Computer Linguistics) at the University of Caen.

Hervé's activities over the past year can be summarised as follows:

- One of the main activities during that period was a formalisation of our approach, and especially the introduction of the notion of Theory refinement. Theory refinement (a.k.a. theory revision or knowledge-base refinement) is the task of modifying an initial imperfect knowledge-base to make it consistent with empirical data. Hervé's method has many similarities with this notion. A bibliographic review on theory refinement

has been achieved. The decision to rewrite the first prototype was taken so that it fits more properly the theory refinement.

- The months of June and July were mainly devoted to the redaction of a technical report describing the learning method.
- After some positive experiments with the Xerox Finite State Tool (XFST), the CASS parser has been abandoned, although it was helpful during the prototyping phase. XFST allows a more convenient formalism to deal with contextual and lexical information. A new version of the learning system is under construction, generating a regular expression grammar using the XFST formalism. The first prototype using the CASS parser is available and was used to evaluate the system.
- Experiments over other non-recursive Phrases (the learning corpus was provided by the ILK group in Tilburg) yield good results. Some problems were encountered in applying the method to the clause structure (in order to achieve the common task). The current learning corpus (WSJ corpus) does not contain adequate information to learn clause structure by using the current method (all the clause boundaries are not marked).
- Online demo of the chunk parser:
<http://www.sfb441.uni-tuebingen.de/~dejean/lcg/chunker.html>
- December 1999: Writing of a paper for the COLING conference.
- February – March 2000: This period was devoted to the study of the influence of the tagset used during the learning. An article was written in March 2000 for the LREC 2000 conference.

Hervé has attended on EACL'99 and CoNLL-99. He has presented his work at the CoNLL-99 (TMR LCG session) and at the LCG Meeting at Dublin. In the framework of the STEEL project, where the chunk parser was used, he participated in the STEEL workshop at the Xerox Research Centre, Grenoble (13-14 September 1999). That was an opportunity to discuss the LCG project with Nicola Cancedda (TMR postdoc).

Three papers by Hervé have been accepted for upcoming international conferences: *How to Evaluate and Compare Tagsets: a Proposal* (LREC 2000), *Theory Refinement and Natural Language Learning* (COLING 2000) and *Applying System Combination to Base Noun Phrase Identification* (COLING 2000, joint work with Tjong Kim Sang, Daelemans, Koeling, Krymolowski, Punyakanok and Roth).

3.7.2 Dale Gerdemann, Coordinator

Dr. Dale Gerdemann (Akademischer Rat, local coordinator), in cooperation with Gertjan van Noord, University of Groningen, has been investigating improvement of Finite State Automata formalisms for Natural Language Processing. Together with Gertjan van Noord, he has presented papers on this topic at EACL'99 and WIA'99. Among his recent publications are *Transducers from rewrite rules with backreferences* (EACL'99, joint work with

Gertjan van Noord) and *Finite Automata with Predicates for NLP* (submitted to ACL2000, joint work with Gertjan van Noord).

3.7.3 Related Tübingen researchers

Prof. Dr. Erhard Hinrichs as director of the research group in computational linguistics at Tübingen, has received three years of funding from the Ministry of Education and Research for morphological and syntactic annotation of text corpora. Tübingen plans to cooperate with TMR-partner Xerox Research Centre Europe, Grenoble in using the XFST tools for syntactic annotation. The project will also provide a testbed for generalising and evaluating the learning techniques developed by Hervé Déjean within TMR.

Sandra Kübler (wiss. Mitarbeiterin) has attended the EACL'99 and CoNLL-99 conferences. She has organised an interest group in Machine Learning and has prepared a German treebank for NP learning. Currently she is working on Learning Lexical Grammar for German.

John Griffiths (wiss. Mitarbeiter) worked on extraction of Cascaded Finite State Grammars from Treebanks (fully parsed German Corpus). The extracted grammars are runnable on the Xerox Finite State Tool (XFST).

3.7.4 Training activities

Hervé Déjean participated in several seminars in Sfs (several talks): Einführung in die Computerlinguistik: Multimediales Lernen und Computerlinguistikcourse and Repräsentation und Annotation linguistischer Korpusdaten. He is a member of the Machine Learning reading group organised by Sandra Kübler. Hervé supervised Klaus Hörmann (Softwarepraktikum). In the months April 1999 – July 1999 and October 1999 – January 2000 he took part in a German course, Internationale Sprachprogramme.

Sandra Kübler visited the Centrum voor Nederlandse Taal en Spraak (Antwerp LCG partner) in December 1999. She received a training on TiMBL.

Rob Koeling (SRI Cambridge) visited the Seminar für Sprachwissenschaft in January 2000 and presented the software Maccent used at SRI Cambridge. Adelina Hild (ISSCO, Geneva) visited Sfs in December 1999.

References

[Adamson and Damper, 1999] Adamson, M. J. and Damper, R. I. (1999). B-RAAM: A Connectionist Model which Develops Holistic Internal Representations of Symbolic Structures. *Connection Science*, 11(1):41–71.

- [Belz, 2000] Belz, A. (2000). Multi-syllable phonotactic modelling. In *Proceedings of SIGPHON 2000: Finite-State Phonology*. Luxembourg.
- [Bouillon et al., 2000] Bouillon, P., Fabre, C., Sébillot, P., and Jacqmin, L. (2000). Apprentissage de ressources lexicales pour l’extension de requêtes. In *T.A.L. 2000*.
- [Bouma, 2000] Bouma, G. (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*, pages 303–310, Seattle. ACL.
- [Buchholz et al., 1999] Buchholz, S., Veenstra, J., and Daelemans, W. (1999). Cascaded grammatical relation assignment. In *Proceedings of EMNLP/VLC-99*. Association for Computational Linguistics.
- [Callan and Palmer-Brown, 1997] Callan, R. and Palmer-Brown, D. (1997). (S)RAAM: An analytical technique for fast and reliable derivation of connectionist symbol structure representations. *Connection Science*, 9(2):139–159.
- [Cancedda and Samuelsson, 2000] Cancedda, N. and Samuelsson, C. (2000). Experiments with corpus-based lfg specialization. In *Proceedings of the 6th Applied Natural Language Processing Conference ANLP-2000*. Seattle, WA, April 29–May 4.
- [Daelemans et al., 1999] Daelemans, W., Buchholz, S., and Veenstra, J. (1999). Memory-based shallow parsing. In *Proceedings of CoNLL-99*. Bergen, Norway.
- [Daelemans et al., 2000] Daelemans, W., Zavrel, J., van der Sloot, K., and van den Bosch, A. (2000). *TiMBL: Tilburg Memory Based Learner, version 3.0, Reference Guide*. ILK Technical Report 00-01.
- [Déjean, 2000a] Déjean, H. (2000a). How to evaluate and compare tagsets: a proposal. In *LREC 2000*. Athens, Greece.
- [Déjean, 2000b] Déjean, H. (2000b). Theory refinement and natural language learning. In *Proceedings of COLING 2000*. Saarbruecken, Germany.
- [Eisele, 1999] Eisele, A. (1999). *Representation and stochastic resolution of ambiguity in constraint-based parsing*. PhD thesis, Stuttgart.
- [Gaussier, 1999] Gaussier, E. (1999). Unsupervised learning of derivational morphology from inflectional lexicons. In *Proceedings of CoNLL-99*. Bergen, Norway.
- [Gerdemann and van Noord, 1999] Gerdemann, D. and van Noord, G. (1999). Transducers from rewrite rules with backreferences. In *Proceedings of EACL’99*. Bergen, Norway.
- [Hammerton, 1998] Hammerton, J. A. (1998). *Exploiting Holistic Computation: An evaluation of the Sequential RAAM*. PhD thesis, School of Computer Science, The University of Birmingham, UK.
- [Hammerton, 1999] Hammerton, J. A. (1999). Holistic Symbol Processing. In Bridge, D., Byrne, R., O’Sullivan, B., Prestwich, S., and Sorensen, H., editors, *Pre-proceedings of the Tenth Irish Conference on Artificial Intelligence and Computer Science, Sept 10-13, University College Cork*, Cork, Ireland. Dept. of Computer Science, University College Cork.
- [Hoste et al., 2000] Hoste, V., Daelemans, W., Tjong Kim Sang, E., and Gillis, S. (2000). Meta-learning for phonemic annotation of corpora. In *Machine Learning: Proceedings of the Seventeenth International Conference (ICML-2000)*. Morgan Kaufmann.
- [James and Miikkulainen, 1995] James, D. L. and Miikkulainen, R. (1995). *SARDNET: A Self-Organizing Feature Map for Sequences*, pages 577–584. MIT Press, Cambridge, MA.
- [Kempe, 1999] Kempe, A. (1999). Experiments in unsupervised entropy-based corpus segmentation. In *Proceedings of CoNLL-99*. Bergen, Norway.

- [Kleiweg and Nerbonne, 2000] Kleiweg, P. and Nerbonne, J. (2000). An fgrep investigation into phonotactics. In Frank van Eynde, I. S. and Schelkens, N., editors, *Computational Linguistics in the Netherlands 1998*, pages 37–50, Amsterdam. Rodopi.
- [Koeling, 1999] Koeling, R. (1999). Using maximum entropy modelling for contextual interpretation of answers. Technical report, Dutch Language and Speech Technology programme.
- [Koeling, 2000] Koeling, R. (2000). Using dialogue information for parsing in a spoken dialogue system. In *Gothenburg Papers in Computational Linguistics*, volume 00-5. University of Gothenburg.
- [Konstantopoulos, 2000] Konstantopoulos, S. (2000). NP chunking using ILP. In Monachesi, P., editor, *CLIN 1999*, pages 109–116, Utrecht. Utrecht Institute of Linguistics OTS.
- [Malouf, 2000] Malouf, R. (2000). The order of prenominal adjectives in natural language generation. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 85–92, Hong Kong. ACL.
- [Malouf et al., 2000] Malouf, R., Carroll, J., and Copestake, A. (2000). Efficient feature structure operations without compilation. *Natural Language Engineering*, 6(1):29–46.
- [Marcus et al., 1993] Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of english: the penn treebank. *Computational Linguistics*, 19(2).
- [Mayberry and Miikkulainen, 1998] Mayberry, M. and Miikkulainen, R. (1998). SARDSRN: A neural-network shift-reduce parser. Technical Report AI98-275, Department of Computer Science, University of Texas at Austin, Texas, US.
- [Nerbonne and Dokter, 1999] Nerbonne, J. and Dokter, D. (1999). An intelligent word-based language learning assistant. *Traitement Automatique des Langues*, 40(1):125–142. Special issue on Multilingual Processing edited by Remi Zajac.
- [Nerbonne et al., 1999a] Nerbonne, J., Heeringa, W., and Kleiweg, P. (1999a). Comparison and classification of dialects. In *Proc. of the 9th Meeting of the European Chapter of the Association for Computational Linguistics*, pages 281–282.
- [Nerbonne et al., 1999b] Nerbonne, J., Heeringa, W., and Kleiweg, P. (1999b). Edit distance and dialect proximity. In Sankoff, D. and Kruskal, J., editors, *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison, 2nd ed.*, pages v–xv. CSLI, Stanford, CA.
- [Osborne, 1999] Osborne, M. (1999). Mdl-based dcg induction for np identification. In *Proceedings of CoNLL-99*. Bergen, Norway.
- [Osborne, 2000] Osborne, M. (2000). Estimation of stochastic attribute-value grammars using an informative sample. In *Proceedings of COLING 2000*. Saarbruecken, Germany.
- [Osborne and Tjong Kim Sang, 1999] Osborne, M. and Tjong Kim Sang, E. (1999). *Proceedings of CoNLL-99*. Bergen, Norway.
- [Ramshaw and Marcus, 1995] Ramshaw, L. A. and Marcus, M. P. (1995). Text chunking using transformation-based learning. In *Proceedings of the Third ACL Workshop on Very Large Corpora*. Cambridge, MA, USA.
- [Reilly, 1999] Reilly, R. (1999). A case study of transient dyslexia. *Brain and Language*, 70:336–346.
- [Reilly, 200x] Reilly, R. (200x). The relationship between object manipulation and language development in broca’s area: A connectionist simulation of greenfield’s hypothesis. *Behavioral and Brain Sciences*. In press.
- [Samuelsson, 2000a] Samuelsson, C. (2000a). A statistical theory of dependency syntax. In *Proceedings of COLING-2000*. ICCL.
- [Samuelsson, 2000b] Samuelsson, C. (2000b). A theory of stochastic grammars. In *Proceedings of NLP-2000*, pages 92–105. Springer Verlag.

- [Sébillot et al., 2000] Sébillot, P., Bouillon, P., Claveau, V., Fabre, C., Jacqmin, L., and Nicolas, T. (2000). Apprentissage en corpus de couples nom-verbe pour la construction d’un lexique génératif. In *JADT2000*. Lausanne.
- [Stoianov and Nerbonne, 2000] Stoianov, I. and Nerbonne, J. (2000). Exploring phonotactics with simple recurrent networks. In Frank van Eynde, I. S. and Schelkens, N., editors, *Computational Linguistics in the Netherlands 1998*, pages 51–68, Amsterdam. Rodopi.
- [Stoianov et al., 1999] Stoianov, I., Nerbonne, J., and Stowe, L. (1999). Connectionist learning to read aloud and comparison to human data. In Hahn, M. and Stoness, S. C., editors, *Proc. of 21st Conf. of Cognitive Science Society*, pages 706–711, Mahwah, New Jersey. Lawrence Erlbaum.
- [Tjong Kim Sang, 2000] Tjong Kim Sang, E. F. (2000). Noun phrase recognition by system combination. In *Proceedings of NAACL 2000*. Seattle, WA.
- [Tjong Kim Sang et al., 2000] Tjong Kim Sang, E. F., Daelemans, W., Déjean, H., Koeling, R., Krymolowski, Y., Punyakanok, V., and Roth, D. (2000). Applying system combination to base noun phrase identification. In *Proceedings of COLING 2000*. Saarbruecken, Germany.
- [Tjong Kim Sang and Nerbonne, 1999] Tjong Kim Sang, E. F. and Nerbonne, J. (1999). Learning simple phonotactics. In Giles, C. L. and Sun, R., editors, *Proceedings of the Workshop on Neural, Symbolic, and Reinforcement Methods for Sequence Processing*. ML2 workshop at IJCAI’99.
- [Tjong Kim Sang and Veenstra, 1999] Tjong Kim Sang, E. F. and Veenstra, J. (1999). Representing text chunks. In *Proceedings of EACL’99*. Association for Computational Linguistics.
- [Veenstra, 1999] Veenstra, J. (1999). Memory-based text chunking. In Fakotakis, N., editor, *Machine learning in human language technology*. workshop at ACAI 99.

A Definition Shared Task 1

In the previous annual report we have presented a general description of three tasks related to learning the syntax of English noun phrases. The idea is that these tasks will be performed by the network participants with their local learning techniques. This will produce interesting material for comparing the strengths and weaknesses of the different machine learning methods when applied to natural language. Meanwhile, a more precise description of the first task has been developed. It can be found at <http://lcg-www.uia.ac.be/con1199/npb/> and we will also present it here.

The first shared task concerns recognising the boundaries of noun phrases (NPs). A lot of work has been done on recognising base NPs, NPs that do not contain other NPs (Ramshaw and Marcus [1995] and follow-up work), but this task aims at finding all NPs. For example:

In (NP early trading) in (NP Hong Kong) (NP Monday) , (NP gold) was quoted at (NP (NP \$ 366.50) (NP an ounce)) .

contains seven NPs of which one contains two embedded NPs.

We propose to use the same training and test data for this task as the popular small baseNP data sets first used by Ramshaw and Marcus [1995] . These contain sections 15-18 of the Wall Street Journal part of the Penn Treebank [Marcus et al., 1993] as training

material and section 20 of the same corpus as test material. The words of these sections and part-of-speech (POS) tags generated by the Brill tagger are freely available to the research community. The url mentioned above contains software for extracting the NP structure from the Penn Treebank. The example sentence converted to the data format used in this task looks like this:

In	IN	*
early	JJ	(*
trading	NN	*)
in	IN	*
Hong	NNP	(*
Kong	NNP	*)
Monday	NNP	(*
,	,	*
gold	NN	(*
was	VBD	*
quoted	VCN	*
at	IN	*
\$	\$	((*
366.50	CD	*)
an	DT	(*
ounce	NN	*)
.	.	*

The task is to use the words and the POS tags to predict the NP structure as well as possible. The performance of the different methods can be evaluated with precision and recall rates. The precision of a method is the percentage of recognised NPs that are correct according to the Treebank annotation. The recall is the percentage of the NPs in the Treebank that are found by the method. The url mentioned above contains a link to `evalb`, a software package written by Satoshi Sekine and Michael Collins which can be used for comparing bracket structures. It computes precision, recall and some other evaluation rates.

Results for this task have been presented at the CoNLL-99 workshop by Tjong Kim Sang (P=90.0% and R=78.4%) and Osborne (P=53.2% and R=68.7%, for a different segment of the Penn Treebank [Osborne, 1999]). In the future, a full parser needs to be applied to the data. An evaluation of its NP bracketing performance would obtain state-of-the-art results which the network participants can attempt to improve on.