

# Learning Computational Grammars

## TMR Project Nr. ERBFMRXCT980237

### 1st Year Report

John Nerbonne and Miles Osborne

September 10, 1999

## Summary

This document describes the first year's progress of the TMR Project *Learning Computational Grammars* (LCG). In brief, despite all sites experiencing delays recruiting staff, LCG now has a full complement of Postdocs, and work is ongoing in areas as diverse as Maximum Entropy, Instance-based Learning, Neural Networks and Genetic Algorithms. In keeping with the original project proposal, all sites are targetting their various learning technologies on the task of learning Noun Phrases in free text. On this note, LCG has created a task description and an attendant training and testing set. The group has produced a series of publications, and has organised a CoNLL meeting, to be held in Bergen at EACL 1999.

## 1 Introduction

After a delayed start, all sites have managed to acquire full-time Postdocs, and in some cases, PhD 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	Symbolic Finite State Methods
Xerox	EBL

The ISSCO group at Geneva will concentrate on complementary work on the linguistic and psycholinguistic subtelties of noun phrases. There have been two group meetings (described in section 2), and apart from familiarisation and exchanging ideas, these have resulted in a common task description (appendix A)

and the creation of attendant training and testing sets. The task description was so-designed to allow all sites to participate in the LCG shared task. A third group meeting will be held in Bergen, June 1999, as part of CoNLL (Computational Natural Language Learning). This year LCG has organised CoNLL, and for the first time in its history, it will have a special focus on Noun Phrase identification in free text (the shared task of all the LCG sites).

The project now has a central web site ([lcg-www.uia.ac.be](http://lcg-www.uia.ac.be)) which summarises the project and perhaps more importantly, acts as publicity.

The rest of this report is as follows. Section 2 describes the two group meetings. The next sections (3 though to 9) are lightly edited site reports. Each report describes LCG-related activities of the Postdocs, PhDs (where relevant) and Coordinators. Following these reports are a list of all project-related publications and finally, in appendix A, the LCG common task definition.

## 2 Group Meetings

### 2.1 TMR-LCG Meeting Groningen June 1998

Local coordinator: John Nerbonne, [nerbonne@let.rug.nl](mailto:nerbonne@let.rug.nl) with Rob Visser, [rob.visser@let.rug.nl](mailto:rob.visser@let.rug.nl) handling finances.

This "introductory meeting" of the TMR project on "Learning Computational Grammars" was structured as an "advanced course" in the 1998 Summer School of Groningen's graduate school in the Behavioral and Cognitive Neurosciences (BCN). This format was chosen to encourage interest from cognitive neuroscientists in attendance at the BCN Summer School, and also to give TMR attendees the reciprocal opportunity to see the rest of BCN work. Evenings and lunches provided opportunities for project planning, and the final session was reserved for this.

The project was pleased to hear an invited guest lecture from Peter Culicover of The Ohio State University, one of the primary contributors to "Learnability Theory". Kenneth Wexler and Peter W. Culicover 1980 *Formal Principles of Language Acquisition*, Cambridge, MIT Press.

- Week 1, morning session, 9:00 - 12:00
  - Monday June 29
    - Opening and Overview (John Nerbonne, Groningen)
    - Phonotactic Learning (Erik Tjong Kim Sang, Groningen and Antwerp)
  - Tuesday June 30: Memory-Based Language Processing
    - Introduction and case study (Walter Daelemans, Antwerp and Tilburg)
    - Implications for Theory of Cognitive Architecture (Gert Durieux, Antwerp)
- Wednesday July 1

- Hard Cases in Syntax (Peter Culicover, Ohio State University)
- Minimal Description Length Learning (Miles Osborne, Cambridge and Groningen)
- Thursday July 2
  - Feature-Based Grammars (Andreas Wagner and Sandra Kuebler, Tbingen)
  - Finite-State Calculus (Dale Gerdemann, Tbingen)
- Friday July 3
  - Connectionist Language Learning (Ronan Reilly, University College Dublin)
  - General Discussion (plenum)

## 2.2 TMR-LCG Meeting Cambridge November 1998

Local coordinators: David Milward david.milward@cam.sri.com and Heather Stewart heather.stewart@cam.sri.com

This meeting featured guest lectures by Stephen Muggleton (York) and Ted Briscoe (Cambridge). Muggleton is one of the primary developers of “Inductive Logic Programming”, and Briscoe is a leading NLP expert on parsing.

- Thursday November 19.
  - 14.00-17.00 Stephen Muggleton, York Tutorial on Inductive Logic Programming
- Friday November 20
  - 09.00 Coffee
  - 09.30 John Nerbonne (Groningen). Welcome and overview of meeting.
  - 09.40 Sylvia Knight (SRI/Cambridge) Decision tree learning
  - 10.10 Rob Koeling (Groningen) MaxEnt and Speech Parsing
  - 10.40 Anja Belz (SRI/Sussex) Discovering FSA by Genetic Search
  - 11.40 Coffee.
  - 12.00 Miles Osborne (Groningen). Evaluation of Grammar Learners.
  - 12.45 Lunch.
  - 14.00 Jorn Veenstra and Walter Daelemans (Antwerp + Tilburg). Memory-based learning applied to NP chunking.
  - 15.00 Ted Briscoe (Cambridge). Automatic acquisition of subcategorization classes from textual corpora.
  - 16.00 Coffee.

- 16.30 Paul John King and Kiril Ivanov Simov (Tuebingen). Clause Fragmentation as a Possible Abduction Process in HPSG-based Language Acquisition.
- Saturday November 21
  - 09.30 Coffee.
  - 10.00 Adelina Ivanova (ISSCO) Research on psychological verbal learning.
  - 10.45 Ivo Stoianov (Groningen). Neural Nets and Phonotactics
  - 11.15 Christer Samuelson, Xerox. (very long title)
  - 11.45 Coffee
  - 12.30 Planning
  - 13.30 Close

### 3 Antwerp

#### Erik Tjong Kim Sang, Postdoc

Before taking the Postdoc position at the Center for Dutch Language and Speech of the University of Antwerp in Belgium, Erik was a PhD student and a part-time teacher at the department of Alfa-informatica of the University of Groningen in The Netherlands (1990-1995) and a full-time teacher at the department of Linguistics of Uppsala University in Sweden (1995-1998). He defended his PhD thesis *Machine Learning of Phonotactics* in October 1998, in Groningen. His thesis supervisor was John Nerbonne. He is a citizen of The Netherlands.

Due to a general project delay the postdoc at Antwerp has started four months later than planned: at August 1, 1998. Since then, most of the time has been spent on research in noun phrase chunking: recognizing basic noun phrases. Erik presented one talk on this topic and wrote one paper about it (joint work with Jorn Veenstra). Apart from this, he defended his thesis, assisted in teaching and graduate student supervision and wrote a task proposal for this project (joint work with Miles Osborne of Groningen). Currently he is preparing an EACL workshop together with Miles Osborne.

Erik Tjong Kim Sang has presented *Automatische verwerving van fonotactische modellen*, in the T&I Colloquia sequence at October 15, 1998 in Tilburg, The Netherlands; *Machine Learning of Phonotactics*, in the Linguistics Colloquium sequence at October 16, 1998 in Groningen, The Netherlands; and *Representing Text Chunks*, at the Computational Linguistics in The Netherlands (CLIN 98) conference, December 11, 1998, Leuven, Belgium.

#### Teaching postdoc

Erik Tjong Kim Sang has taught in the following undergraduate courses in Belgium: Computational Linguistics, Antwerp, November 1998 (four hours);

Language Technology, Antwerp, March 1999 (two hours); and Computational Linguistics, Ieper, March 1999 (three hours).

#### **Miscellaneous postdoc**

Erik Tjong Kim Sang has visited the following conferences and meetings: Benelearn 98, Wageningen, The Netherlands, October 1998; TMR-LCG meeting, Cambridge, UK, November 1998; and CLIN 98, Leuven, Belgium, December 1998. He has defended his PhD thesis, in Groningen, The Netherlands in October 1998.

Together with Miles Osborne he is organising the CoNLL workshop which will take place in June 1999. These two have also written a document *TMR-LCG Core Research Tasks* which was distributed to project participants by e-mail in December 1998.

Erik Tjong Kim Sang has been spending one working day in the week at the TMR-LCG partner Tilburg from December 1998.

#### **Walter Daelemans, Coordinator**

A number of other current projects of CNTS involve related research on Machine Learning of Natural Language. In the LINGUADUCT project (a Belgian government sponsored project in cooperation with the University of Leuven), machine learning of information extraction rules is investigated, and the use of learning to automatically model pronunciation differences between regional variants of Dutch for applications in speech synthesis (Veronique Hoste). Guy De Pauw started a Belgian NSF funded project on modeling the evolution of language with ML techniques. Within the field of computational psycholinguistics, memory-based learning is used as a simulation tool for modeling child language acquisition (Masja Kempen, research in cooperation with Utrecht University), and as a model of adult language processing (in a new project in cooperation with the psycholinguistics group of UFSIA in Antwerp; Helena Taelman).

For teaching: see Erik's report. No explicit training for postdocs.

The University of Antwerp has a doctoral education programme where pre-docs can get credits by following courses, going to and participating in scientific events, and by teaching. No such program exists for postdocs.

#### **Others**

At this site two people are involved: the postdoc Erik Tjong Kim Sang and the local project coordinator Walter Daelemans. There are strong connections between this site and the ILK group in Tilburg. From that site two more people are involved: Jorn Veenstra and Sabine Buchholz.

## **4 Dublin**

#### **James Hammerton, Postdoc**

This report summarises the progress James have made on this project since starting the postdoctoral fellowship here at University College Dublin (UCD), on the 11th January 1999. Due to the need to find a place to live, and being required to make some modifications to my PhD thesis, the work on the project did not start properly until late February.

Thus far the work has involved evaluating some neural network technologies for use with the LCG tasks. He has decided to investigate whether a combination of Mayberry & Miikkulainen's SARDSRN architecture (Mayberry & Miikkulainen 1998) and Callan & Palmer-Brown's Simplified RAAM or (S)RAAM (Callan & Palmer-Brown 1997) can be used for the LCG tasks:

- The SARDSRN seems to be a promising connectionist architecture for parsing, offering considerably improved performance over the Simple Recurrent Network (SRN). It combines the SRN with a variation of the Self-Organising Map architecture (Kohonen 1990) known as SARDNET (Self-organising Activation-Retention Decay NETwork). SARDNET modifies the SOM so that it can map sequences, and this map is used as input to the SRN to help it retain information over longer time periods. James' experiments with SARDSRN, and correspondence with Marshall Mayberry over how best to implement it confirm that it offers better performance for natural language tasks than the SRN, and thus may be able to scale to a large-scale task such as the LCG tasks. However in order to use it for this purpose some method of representing output needs to be devised.
- The (S)RAAM is an analytical technique for developing connectionist representations of compositional structures, such as trees or lists. It is a derivative of Pollack's Recursive Auto-Associative Memory (RAAM) (Pollack 1990), but offers improved training times (taking minutes as opposed to hours or even days), greater reliability and guaranteed generalisation properties (it generalises to linear combinations of the data it sees in the training set). However it does generate larger representations than the RAAM. James has obtained C++ code implementing the (S)RAAM from a research student of Robert Callan. He is currently investigating whether the representations generated from a set of trees from the Wall Street Journal corpus will be of a practical size to use with the SARDSRN. Preliminary results suggest it may generate extremely large vectors (500 elements or more) however I have yet to verify whether this is due to some bugs in the code or whether it is a property of (S)RAAMs. Such large vectors may make training the SARDSRN prohibitively expensive.

Thus far, the SARDSRN looks promising but the (S)RAAM is less so, though more work needs to be done to see if this is definitely the case. Should the (S)RAAM turn out not to be so promising there are other representational techniques that can be tried, such as Holographic Reduced Representations (Plate 1991), the RAAM itself (Pollack 1990) or another recent improvement on the RAAM, known as the Bi-coding RAAM (Adamson & Damper 1999). The motivation for using RAAM-style techniques such as those mentioned above is

that they support the holistic processing of symbol structures, a phenomenon currently unique to connectionist systems, but which has yet to be exploited in a large-scale task such as the LCG tasks.

In addition to this work, James has also given a talk on my work as part of the Artificial Intelligence Seminar series held in the department and he has had discussions of my work with other members of the department.

Together with Ronan Reilly, James is co-supervising an undergraduate student, Paul Heffernan, on a summer internship at the department. The internship involves investigating whether a modification to the training of a neural network known as the Sequential RAAM (SRAAM) can improve its performance as a technique for representing compositional structures in neural nets. The student will have to extend the PDP++ neural network simulator to include the modified training scheme and perform experiments to evaluate whether the modifications do in fact help with both training the SRAAM and the performance of the trained SRAAM. This essentially ties up a loose end from my PhD work and gives the student experience in research with neural nets and a modest amount of C++ programming.

James recently completed a PhD at the School of Computer Science, Birmingham University, UK. The title of the thesis is “Exploiting Holistic Computation: An evaluation of the Sequential RAAM”, and he was supervised by Peter Hancox.

### **Ronan Reilly, Coordinator**

Ronan has presented a paper related to the project to the ICANN'98 conference (Reilly 1998).

### **Others**

Arthur Cater and Julie Berndsen have also taken part in project-related discussions. Ivelin Stoianov of Groningen visited for two weeks in March.

## **5 Geneva**

### **Adelina Ivanova, Postdoc**

One of the challenges for parsing nominal phrases is posed by deverbal nominalizations. Automatic parsing has so far focussed on identifying complements and specifiers in noun phrases headed by non-derived nominals, an area where research efforts have been somewhat successful. Work on automatic processing of nominalizations have so far been stalled by the many questions regarding their grammatical structure with which linguists are still faced (cf. Spencer, 1995). Understanding the syntactic properties of nominalizations and the underlying semantic features means not only more efficient parsing of a variety of syntactic structures, but is also requisite for future applications in the area of machine translation. Research in human language processing has suggested that translation of some types of nominalizations can be problematic, which was attributed

to the fact that similar semantic features are projected on the morphosyntactic level in a language-specific way (Ivanova, 1998).

At the this stage of the project ISSCO's contribution is targeted at these issues with a view to developing an annotation standard for automatic parsing for these grammatical constructions. The work so far involved assessing available typologies of nominalizations. Recent typologies offered by theoretical linguists have been studied in detail. Koptjevskaja-Tamm (1993) offers a semantic typology, which although based on a wide-ranging survey of genetically diverse languages, is not exhaustive. Bauer (1983) provides a list of meanings for nominalizations (as many as 13 distinct meanings), which need to be analysed in the light of more recent classification attempts (cf. Koptjevskaja-Tamm, 1993). Grimshaw (1990) attempts to relate the semantics of the nominalization to its realisation on the morphosyntactic level as argument structure. Although the latter represents a new approach to the typology of the nominalization, it encounters problems. For instance, problems with the distinctions between modifiers and complements offered by Grimshaw leave open the question of how to treat prenominal possessives and adjectives. Furthermore, as Spencer (1995) points out, some of the Grimshaw's claims, based on the analysis of English nominalization alone, are not supported by linguistic data from other languages.

In this connection it is proposed that the work on the project will proceed by investigating a large database from diverse languages, focussing initially on German and English. Furthermore, connections have been established with the Department of Linguistics, University of Essex, where on-going research investigates the aspectual structure of nominalizations in Russian. Thus, it is hoped that from the analyses of diverse languages certain cross-linguistic similarities will emerge which will allow us to explore further the semantic features accounting for the syntactic behaviour of the nominalization in order to develop a typology as a basis for workable annotation for this type of nominal constructions. It should also be mentioned that the multilingual data which ISSCO is in the processes of preparing should be useful for the other partners on the project.

**Susan Armstrong, Coordinator**

### **Others**

John Nerbonne visited in March to discuss staffing and the Geneva contribution to LCG.

## **6 Groningen**

**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 30<sup>th</sup> of September 1998 and started the current position on the 1<sup>st</sup> of October 1998.

LCG related research activities are summarised as follows:

- Adaptation of the Sparkle DCG learner for NP identification. This task was straightforward, given the fact that the learner acquired NP rules already. The main change was allowing it to be trained on parsed corpora annotated with NP information.
- Induction of DCGs modelled as random fields. 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. As a by-product, random-field modelling should lead to competitive parse selection results.
- Organisation of CoNLL99 ([lcg-www.uia.ac.be/conll99](http://lcg-www.uia.ac.be/conll99)) in conjunction with Erik Tjong Kim Sang (Antwerp).
- Preparation of LCG task definition and evaluation framework, again in conjunction with Erik Tjong Kim Sang.
- Talk given on evaluating NP learners, SRI, UK, November 1998.

Non-LCG academic activities are summarised below:

- Supervision of Tony Mullen (PhD student), working on random field-based parse selection.

- Preparation and delivery of a course on Statistical Natural Language Processing here in Groningen. This is available at <http://odur.let.rug.nl/osborne/>

### **Stasinos Konstantopoulos, PhD student**

Stasinos is using Progol (an Inductive Logic Programming system developed in York ([www.cs.york.ac.uk/stephen/progol.html](http://www.cs.york.ac.uk/stephen/progol.html)) to induce a DCG for Noun Phrases in English from the Wall Street Journal annotated corpus (part of the Uni of Pennsylvania Treebank II project).

He is currently experimenting with positive only training data and very simple or no (linguistic) features, and trying to establish the exact capabilities of the Progol system to learn features. The goal is to estimate the feasibility as well as the importance of developing an ILP system that will be able to learn a DCG grammar that will employ features to collapse multiple rules while at the same time constraining against over-generalisation. The learning task is to be restricted to the domain of noun phrase syntax of Germanic languages, although care must be taken to allow extendability to a full grammar.

### **John Nerbonne, Coordinator**

John Nerbonne supervised the PhD thesis of Erik Tjong Kim Sang, *Machine Learning of Phonotactics*, PhD thesis, University of Groningen, 1998, defended in Oct. 1998. Erik is a postdoc in LCG at the Antwerp site.

John Nerbonne continued work on phonotactic learning in neural networks with Ivelin Stoianov, presenting “Exploring Phonotactics with Simple Recurrent Networks” to appear in: Frank van Eynde, Ineke Schuurman and Ness Schelkens (eds.) Proc. of Computational Linguistics in the Netherlands 1998, and also with Peter Kleiweg, presenting “An FGREP Investigation into Phonotactics”, to appear in the same volume.

He also investigated the classification of dialects using unsupervised learning (clustering), and visited the Geneva site in March.

### **Others**

Ivelin Stoianov (PhD student supervised by John Nerbonne) made the presentation with Nerbonne noted above, and he also visited the Dublin LCG site for the period of 1-15 March 1999. There he studied a novel learning algorithm for recurrent Neural Networks called “Long Short Term Memory”. Unlike existing learning algorithms, such as Error Back-Propagation or Back-Propagation, this new approach claims to have better performance when dealing with long-distance relations (in time). During his visit, Stoianov presented a talk on learning lexical grammar (phonotactics) with neural networks.

Gosse Bouma (lecturer) is on the *Language, Learning and Logic* programme committee <http://www.cs.york.ac.uk/mlg/111/workshop>.

## 7 SRI

### Anja Belz, Postdoc

Background.

M.A. Technical Translation (1994, University of Westminster), M.Sc. Computing Science (1995, Imperial College).

Research for the degree of D.Phil. in Cognitive Science and Artificial Intelligence since Oct 1995 (University of Sussex), Supervisor: Prof. Gerald Gazdar, Subject: the development of a formal method for phonotactic description and of practical learning techniques for the automatic construction of such descriptions. Main learning method investigated: a specially developed genetic algorithm for the automatic construction of finite-state automata that generalise over given phonological data samples.

Other research interests include morphology, general automata theory, neural networks, comparison of natural language learning methodologies, and speech recognition.

#### Summary of Project Activities to May 1999

- Employed part-time (1 day per week) January - April 1999, full-time since May 4th 1999.
- Presented a paper at the Cambridge LCG project meeting (“Discovering FSAs by genetic search”).
- Literature review of existing genetic algorithm and evolutionary techniques for finite-state and context-free grammar induction, to be presented in a technical report in the near future.
- Design of a webpage on genetic algorithms to be included in the official LCG website, as part of a collection of webpages on the different learning methods used by LCG researchers.
- Adaptation of existing GA for constructing phonotactic finite-state automata to the task of learning the structure of NPs under a finite-state constraint.
- Design and implementation of a series of experiments learning NP structure by generalising over data sets of tagged NPs.

#### Research Plans to April 2001

- **May - October 1999** Investigation of different representational schemes for context-free grammars. Design of further experiments involving NP structure. Extensive testing of the resulting GAs on real tasks e.g. as a module of SRI's Highlight System. (See [www.cam.sri.com/html/highlight\\_demo.html](http://www.cam.sri.com/html/highlight_demo.html)).

- **October 1999 - April 2000** Investigation of different approaches to integrating background knowledge into GA learning methods, carrying out further tests on real tasks. At the end of first year of research, review of the overall suitability of genetic algorithms to the shared LCG tasks and presentation of the results in a technical report.
- **May 2000 - April 2001** Investigation of alternative learning methods (especially ILP and ME), and of how elements of these may be combined with a GA for the task of learning NP boundaries and structure. Continue to refine and run experiments, focus on analysis of results.

## **Rob Koeling, PhD student**

**Background:** Bachelors degree in Computing Science; Masters degree in Computational Linguistics. Worked previously on the grammar of a natural language processing module of a spoken dialogue system. PhD thesis work (Groningen) is on using contextual (dialogue) knowledge to improve wordgraph parsing. He previously looked at knowledge based approaches, currently Rob is investigating the use of statistical (maximum entropy) models to exploit information in system questions for parsing user utterances.

- Talk: "MaxEnt and Speech Parsing" TMR-LCG Meeting Cambridge November 1998
- Talk: "A Maximum Entropy model for modelling context in a spoken dialogue system" Computational Linguistics in the Netherlands. Katholieke Universiteit Leuven, December 11, 1998.

**Research Plans:** In the period April 1999 to May 2000 Rob plans to investigate the possibilities of applying the Maximum Entropy modelling techniques to the research tasks defined for the project.

## **David Milward, Coordinator**

David Milward is the local project coordinator. He has particular interests in applying the results of the project to improve the parsing components of text processing systems used at SRI. This year he attended the LCG kickoff meeting in Groningen and hosted the Cambridge meeting. In the project he has responsibility for overseeing Anja Belz's work for the LCG project.

## **Others**

### **1. Richard Sharman, Director, SRI**

Richard Sharman takes an active interest in Maximum Entropy techniques and will be involved in overseeing Rob Koeling's work for the LCG project using ME techniques.

2. **Stephen Pulman, Principal Scientist (SRI) and Reader (University of Cambridge Computer Laboratory)**

Stephen Pulman played a central role in the LCG project during its preparatory stages, particularly in planning the Cambridge part of the project, and has attended meetings and related activities (e.g. those involving ILP). He intends to remain actively involved, relating the work of local full-time LCG researchers to the work of researchers at the Computer Laboratory (especially David Abensour and Sylvia Knight) and to his own interests in ILP. SRI is a member of the EC ILP2 end-users club, and Stephen Pulman has been helping to keep that project and LCG in touch.

3. **Ted Briscoe, Lecturer, University of Cambridge Computer Laboratory**

Ted Briscoe presented a talk at the Cambridge LCG project meeting, (“Automatic acquisition of subcategorization classes from textual corpora”). He is currently doing mark up of NP-internal structure in the Susanne Corpus test set that is going to be used for evaluation purposes by project members.

4. **Sylvia Knight, Part-time researcher (SRI) and Doctoral Researcher and Tutor (University of Cambridge Computer Laboratory)**

Sylvia Knight presented a talk at the Cambridge LCG project meeting (“Decision tree learning”), and will remain involved on a small scale through discussions and meetings.

Collaboration between the local full-time LCG researchers and associated project members, in particular at the **University of Cambridge Computer Laboratory**, forms an important part of the Cambridge LCG activities. A special interest group in Inductive Logic Programming (Stephen Muggleton, Stephen Pulman et al. ), has been very active organising projects, seminars and meetings. The plan is to extend collaboration between ILP projects and LCG further, in the form of regular meetings and discussion groups.

There are plans for local and general profile-raising activities, such as regular talks at UK departments and research institutes, visits to LCG partners, and a local webpage presenting project-related information including regular updates, project reports and technical papers.

## 8 Tuebingen

**Hervé Déjean, Postdoc**

## Realization

The first three months were devoted to the achievement of a base Noun phrase chunker, considered as the first stage of the task 1 (recognizing NP structures) defined in (Osborne & Sang 1998). This program automatically learns finite state grammar from bracketed corpus. The implementation of this chunker is now complete, and the evaluation done. The results (better than those obtained by Ramshaw and Marcus) validates the hypothesis that the finite state framework, widely used in handcrafted approaches, seems also suitable for a learning task. The approach used in this learning algorithm is the generation of a finite state grammar that, when used with the Cass system, parses English Noun Phrases. The data and the software evaluation used are those described at the TMR-LCG site page: <http://lcg-www.uia.ac.be/lcg/resources>.

The work described in (Déjean 1998) offering a method for chunk extraction from raw texts, was used as starting point. The grammar learning relies on a grammar scheme variant of X-bar theory. It is currently applied to the chunking task, but, like the genuine X-bar scheme, it can be applied on other structures (verbal phrase, simplex clause), and thus will be used during the next stage of the task 1 (recognizing all NP structures). The use of this scheme partially allows us to achieve, for baseNPs, the second task of the project: recognizing the internal structure. The grammar generated uses the Cass system developed at the University of Tübingen by Steven Abney. But, experience shows us that this system is not suitable for the integration of lexical information, which would allow improvement of our results. At the current stage of our work, the lexical integration needs postprocessing.

## Future works

During the coming months, our purpose is to deal with two main directions:

As said above, the Cass system offers a weak formalism to take into account lexical level. Two options are offered to us: the first is to use another tool such as the Xerox Finite State Tools package. The second option is to improve the Cass system by including new features.

The second direction will be to complete the task 1 by extending the learning to all NP structures. This work will be in relation with the work at the University of Tübingen: one result of work done was the creation of several handed-coded tree banks of German conversations. These tree banks are now being used in project A1 of the SFB 441 as a source of grammar rules. The goal of this work is to extract from these banks cascaded finite state grammars for use with the Xerox Finite State Tools package as well as with the Cass system.

## Particulars

- **Thesis:** Concepts and Algorithms to Discover Natural Language Syntactic Structures. Supervisor: Prof. Khaldoun Zreik, Dr. Jacques Vergne, GREYC - CNRS 6072, University of Caen.

- **Last position:** Attaché temporaire d'enseignement et de recherche, (one year research and teaching position), Computational Department, University of Caen, France.

## **Dale Gerdemann, Coordinator**

### **Others**

- Erhard Hinrichs (German corpus)
- Sandra Kübler (German corpus)
- John Griffith (Grammar extraction, finite-state tool)

### **Related work at the site**

That work is related with several projects done at the University of Tübingen.

- In the framework of the STEEL project (Copernicus program), the chunker developed will be used for the Noun Phrases extraction. The learning algorithm will be applied on the specific corpus used in that project.
- The SFS is developing a German Shallow Parser in collaboration with Xerox (Grenoble Center). The learning algorithm will be applied on the treebank corpus provided by the Verbmobil project in order to extract grammar rules.

### **Training**

The progress of the work was presented in the course “Einführung in die Computerlinguistik: Multimediales Lernen und Computerlinguistikcourse”, Seminar für Sprachwissenschaft. The topics were a general presentation of the chunking and shallow parsing problem, and a presentation of the different learning methods used by the TMR participants.

## **9 Xerox**

### **Nicola Cancedda, Postdoc**

Nicola Cancedda joined the LCG project as a TMR postdoc on March 1, 1999. His research interests so far were mostly focused on Information Extraction and Natural Language Generation. He is currently completing his PhD program in computer engineering at the University of Rome “La Sapienza”, where he also obtained his MEng, with a thesis on the generation of text from the output of Information Extraction systems. As a part of his PhD program, he spent six months at the Artificial Intelligence Center of SRI International, in Menlo Park, CA. All along his Phd program, Nicola Cancedda also consulted for Finsiel SpA, where he participated in a joint project with the Natural Language Processing

group of the “Istituto per la Ricerca Scientifica e Tecnologica (IRST)” of Trento - Italy. The project lead to a prototype system for generating hypertextual descriptions of SADT conceptual models of information systems.

The established goal for the LCG project at XRCE is the application of Explanation Based Learning (EBL) to specialize Lexical Functional Grammars on particular domains. The current work focuses on the application of EBL to Lexical Functional Grammars in the form required by the Xerox Linguistic Environment.

Active work on the project began, at the XRCE site, in the middle of March 1999. The activities conducted so far include:

- **Study of previous research on EBL for parsing.** A bibliographical research was conducted in order to survey previous research on the application of EBL to parsing.
- **Assessment of available resources.** Two treebanks previously generated by means of LFGs for French and English developed at Xerox were found available. Both of them -composed of 1000 parses of sentences each- were inspected to verify their suitability for applying EBL. Moreover, part of the time was devoted to assessing the impact on the learning process of the primitives that the XLE provides to the grammar writer.
- **Definition of the experimental framework.** The problem of applying EBL to LFGs was decomposed, and the overall framework for the experiments to be carried out was defined.
- **Preliminary problem: joining grammar rules.** Whereas the problem of joining rules is trivial for CFGs, it is not for LFGs, as the functional schemas on suppressed nodes need to be absorbed by surviving nodes. This problem was solved as a suitable algorithm was found.
- **Preliminary problem: relating a parse to the rules used to derive it.** The problem of figuring out what rule was used to expand each nonterminal in a parse-tree is not trivial for LFGs, both because of the presence of functional schemas and because rules allow regular expressions on symbols in their RHSs. The overall structure of an algorithm for solving this problem was sketched.

**Christer Samuelson, Coordinator**

**Others**

## References

- Adamson, M. J. & R. I. Damper (1999), ‘B-RAAM: A Connectionist Model which Develops Holistic Internal Representations of Symbolic Structures’, *Connection Science* **11**(1), 41–71.
- Belz, Anja (1998*a*), An approach to the automatic acquisition of phonotactic constraints, *in* ‘Proceedings of SIGPHON ’98: The Computation of Phonological Constraints’, , University of Montreal, Canada, pp. 35–44.
- Belz, Anja (1998*b*), Discovering phonotactic finite-state automata by genetic search, *in* ‘Proceedings of COLING-ACL ’98’, Morgan Kaufman, pp. 1472–1474.
- Belz, Anja & Berkan Eskikaya (n.d.), A genetic algorithm for finite-state automaton induction. *CSRP 487*, School of Cognitive and Computing Sciences, University of Sussex.
- Buchholz, Sabine (1998*a*), Distinguishing complements from adjuncts using memory-based learning, *in* ‘Proceedings of the ESSLLI-98 Workshop on Automated Acquisition of Syntax and Parsing’.
- Buchholz, Sabine (1998*b*), Learning Subcategorization, *in* H. Strating & J. Veenstra, eds, ‘Proceedings of the CLS opening Academic Year ’98 ’99, Center for Language Studies’, Tilburg, The Netherlands.
- Buchholz, Sabine (1998*c*), Unsupervised learning of subcategorisation information and its application in a parsing subtask, *in* H. L. Poutre & H. van den Herik, eds, ‘Proceedings of the Tenth Netherlands/Belgium Conference on Artificial Intelligence (NAIC’98)’, CWI, Amsterdam, The Netherlands.
- Callan, R.E. & D. Palmer-Brown (1997), ‘(S)RAAM: An analytical technique for fast and reliable derivation of connectionist symbol structure representations’, *Connection Science* **9**(2), 139–159.
- Daelemans, Walter (1998), Toward an exemplar-based computational model for cognitive grammar, *in* J. van der Auwera, F. Durieux & L. Lejeune, eds, ‘English as a Human Language. To honour Louis Goossens’, LINCOM Europa, München, pp. 73–82.
- Déjean, Hervé (1998), Concepts et algorithmes pour la découverte des structures syntaxiques des langues, PhD thesis, Université de Caen.
- Gerdemann, Dale & Gertjan van Noord (1999*a*), Transducers from rewrite rules with backreferences, *in* ‘EACL’99’, Bergen.
- Gerdemann, Dale & Gertjan van Noord (1999*b*), ‘Transducers from rewrite rules with backreferences’, EACL99.

- Hans van Halteren, Jakub Zavrel & Walter Daelemans (1998), Improving Data Driven Wordclass Tagging by System Combination, *in* ‘Proceedings of the 36th annual meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics’, Montreal, Canada, pp. 491–497.
- Karttunen, Lauri, Tamás Gaál & André Kempe (1997), *Xerox Finite-State Tool*, Xerox Research Centre Europe, Grenoble.
- Koeling, Rob (n.d.), Using maximum entropy modelling for contextual interpretation of answers. TST Technical Report (forthcoming).
- Kohonen, T. (1990), ‘The self-organising map’, *Proceedings of the IEEE* **78**(9), 1464–1480.
- Mayberry, M. & R. Miikkulainen (1998), SARDSRN: A neural-network shift-reduce parser, Technical Report AI98-275, Department of Computer Science, University of Texas at Austin, Texas, US.
- Osborne, Miles (1999a), *DCG induction using MDL and Parsed Corpora*, *in* J. Cussens, ed., ‘Language Learning in Logic’, Bled, Slovenia, pp. ??–?? Workshop held in conjunction with ICML99.
- Osborne, Miles (1999b), MDL-based DCG Induction for NP Identification, *in* M. Osborne & E. F. T. K. Sang, eds, ‘Proceedings of CoNLL99’, EACL, Bergen, Oslo, pp. xx–yy.
- Osborne, Miles & Erik F. Tjong Kim Sang, eds (1999), *CoNLL99: Computational Language Learning*, EACL, Bergen, Norway.
- Osborne, Miles & Erik Tjong Kim Sang (1998), *TMR-LCG Research Tasks*.
- Plate, T. A. (1991), Holographic Reduced Representations: Convolution algebra for compositional distributed representations, *in* J. Mylopoulos & R. Reiter, eds, ‘Proceedings of the 12th International Joint Conference on Artificial Intelligence, Sydney, Australia, August 1991’, Morgan Kaufman, San Mateo, CA, pp. 30–35. Reprinted in Mehra P. and Wah B.W. (editors). *Artificial Neural Networks: Concepts and Theory*, Los Alamitos, CA, IEEE Computer Society Press 1992.
- Pollack, J. B. (1990), ‘Recursive distributed representations’, *Artificial Intelligence* **46**(1–2), 77–105.
- Reilly, R.G. (1998), Enriched lexical representations, large corpora and the performance of srns, *in* L. Niklasson, M. Bodèn & T. Zemke, eds, ‘Proceedings of ICANN’98, Skovde, Sweden’, pp. 405–410.
- Stegmann, R., H. Schulz & E. Hinrichs (1998), Stylebook for the german treebank in verbmobil, Technical report, Seminar für Sprachwissenschaft, Universität Tübingen.

- Tjong Kim Sang, Erik F. (1998), Machine Learning of Phonotactics, PhD thesis, University of Groningen, The Netherlands.
- Tjong Kim Sang, Erik F. & John Nerbonne (n.d.), 'Learning Simple Phonotactics', IJCAI-99 Workshop on Neural, Symbolic, and Reinforcement Methods for Sequence Learning.
- Tjong Kim Sang, Erik F. & Jorn Veenstra (1999), 'Representing Text Chunks', EACL99.
- Veenstra, Jorn (n.d.), 'Fast np chunking using memory-based learning techniques'. F. Verdenius and W. van den Broek (eds).
- Walter Daelemans, Jakub Zavrel & Antal van den Bosch (1999a), 'Forgetting exceptions is harmful in language learning', *Machine Learning* (34), 11–41.
- Walter Daelemans, Jakub Zavrel, Ko van der Sloot & Antal van den Bosch (1999b), TiMBL: Tilburg Memory Based Learner, version 2.0, Reference Guide, Technical Report Technical Report 99-01, ILK, Tilburg.

## A Common Task Description

### A.1 Introduction

This is a proposal description for the core research tasks in the TMR Network Learning Computational Grammars. It gives an outline of research tasks, the text corpora, scoring metrics and task performance evaluation in this network. Participants are expected to perform at least one of these tasks with the mentioned corpora. However, they are free to explore other tasks or other corpora within the network but performing the core tasks will enable them to compare their results with the results of the other network participants.

### A.2 Tasks

The project proposal states that we will apply machine learning techniques for learning NP syntax. We suggest that we focus upon the three tasks outlined in the TMR homepage [Ner98]:

- Annotating sentences with parentheses marking NP boundaries.
- Recognising the internal structure of recognised NPs (ie drawing parse trees for NPs).
- Retrieving grammatical relations relevant to recognised NPs.

The first task concerns recognising \*all\* NP boundaries in sentences in which words have been marked up with part-of-speech tags. There have been a number of publications on NP chunking, the recognition of so-called baseNPs (for example [RM95]). These studies do not consider NPs which contain other NPs. However, in this network we propose to define NP boundary recognition as the recognition of all NP boundaries.

The parts of the sentence that should be regarded as NPs are defined in the corpora used by the participants (see the Corpora section for a more detailed description).

In the second task amounts to finding the syntactic structures within NPs. The tree labels that need to be discovered here will be dependent of the corpora that is used in this task. The input of this task contains words, part-of-speech tags and NP boundaries.

The third task causes some practical problems. The two corpora which we want to propose contain sufficient annotation detail for the first two tasks but not for the third task. Ted Briscoe has given some input on this front -see the A.5. Note also that Sparkle have said that we have access to 500 SUSANNE sentences that have been partially marked-up with grammatical relation information. These relations are extra to any material found in SUSANNE. However, this annotation is partial, and we would have to complete the task.

The list of tasks does not contain a task for determining head noun, modifiers and determiner. [Ner98] mentions such a task. However, we believe that head

nouns, modifiers and determiners can be determined trivially from the parse tree that will be obtained in the second task.

Network participants are allowed to work on different tasks but ought to put some work in at least one of the three tasks mentioned here.

### A.3 Corpora

We suggest using two corpora: the Wall Street Journal (WSJ) corpus and the SUSANNE corpus. Both corpora are available to the research community. WSJ is large and it is used on a large scale in our field. However, there are some concerns with respect its annotation quality and therefore we suggest using the smaller SUSANNE corpus for cross-comparison. All things being equal, improved performance on WSJ should also mean improved SUSANNE performance. What is more likely however are high figures for WSJ and lower ones for SUSANNE.

WSJ is a part of the Penn Treebank and can be ordered from the Linguistic Data Consortium: <http://www.ldc.upenn.edu/> . Tree structures that are dominated by NP or WHNP should be considered as NPs. In some cases these tags contain some extra attributes separated by hyphens, like in NP-SUBJ-1. These extra attributes should not be taken in consideration when deciding whether a structure is an NP or not.

SUSANNE is a public domain corpus. You can download it from the Oxford Text Archive: <ftp://ota.ox.ac.uk/pub/ota/public/susanne/> . Here local trees dominated by symbols N, Nns, Np or Ns should be regarded as NPs. These tags can contain an extra hyphen or some extra attributes separated by colons, like in Np:s. Again these extra characters should not be taken in consideration when deciding whether a structure is an NP or not.

Participants can propose using other corpora, in particular corpora containing non-English text. These corpora should be of a quality comparable, if not better than SUSANNE.

### A.4 Metrics and evaluation

An evaluation of results is not trivial because we need to compare parse trees with each other. Here we cannot rely on exact matches only because that would assign low evaluation scores to most systems. This would prevent a spread of results, and so make a clear comparison harder.

A discussion of parser evaluation metrics can be found in [CBS97]. Based on this document we suggest the following evaluation schemes:

- task 1: unlabelled GEIG
- task 2: labelled GEIG
- task 3: Sparkle GR

The GEIG evaluation scheme (Grammar Evaluation Interest Group) involves three parse comparison rates: precision, recall and crossing brackets [HABFG91].

Precision is the number of correctly recognised constituents divided by the number of constituents found, recall is the number of correctly recognised constituents divided by the number of constituents in the corpus.

The crossing brackets rate contains the number of recognised constituents that violates actual corpus constituent boundaries. The difference between the two GEIG variants is that labelled GEIG only considers constituents to be correct when they have a correct label. For unlabeled GEIG the value of the labels is not important.

In the Sparkle GR evaluation scheme parse trees are interpreted as a collection of grammatical relations [CBS97]. These include for example the relations between heads, modifiers and arguments. This scheme includes two evaluation rates: precision and recall. Apart from the fact that they consider relation scores rather than constituent scores they are computed in the same way as in GEIG. The correctness specification for relations is a little bit more relaxed than identity (for details see [CBS97]).

The (revised) evaluation scheme would consist of sites taking test sets (unannotated sentences) and then annotating these test sets in a manner relevant to the given task. The annotated test sets would then be evaluated using standard software (eg Parseval, EVALB etc). This evaluation would preferably be performed by a central site.

## A.5 Grammatical Relations

Ted Briscoe made the following suggestions regarding Grammatical Relations (GRs):

The Sparkle GR scheme is described at:

<http://www.ilc.pi.cnr.it/sparkle/wp1-prefinal>

The relation mod(ifier) is used to describe both adjectival modification and ofPP arguments of N heads. The definition is:

mod(type,head,dependent)

eg

mod(X,flag,red) a red flag

mod(of,gift,book) the gift of a book

mod(of,examination,patient) the examination of the patient

The specialisations of mod x/cmod, ncmmod (with same args) divide mods into clausal (open controller/controlled within) and non-clausal mods. This stuff is taken from the Eagles standard and was developed by Antonio Sanfilippo. I haven't checked back to see whether the following are dealt with in this much bigger document which is also available on the web – a quick fix might be to ask Antonio!

Probably for a better treatment of NPs it wld be worth distinguishing adjuncts from arguments to deverbal nouns in annotation,

but, of course, few if any of the parsers would be able to distinguish them so this shld be done by specialising mod with a fourth subtype eg. amod:

amod(of,examination,patient) the examination of the patient

so everyone else can default back to mod. Possessive shld probably also be a type of mod:

John's car  
mod(poss, car, John)

and this raises the question of whether you want this to be more specialised in:

John's examination of Mary

eg.

amod(poss, examination, John)

This at least begins to represent the fact that possessives and of-PPs with deverbals tend to be interpreted as subj/agt and obj/pat, respectively, but doesn't bind you to this mapping – 'John's examination' can be the one he received.

Beyond this I personally don't see much point in diverging from what is given. You already have quite a complex task, say, deriving the following relations from this eg:

The village over the hill by the main road  
mod(over, village, hill)  
mod(by, village, road)

That is, you have to be able to do the correct attachments of pre/post modifiers to derive the correct relations, so the scheme abstracts appropriately from other details of syntactic description employed by different parsers.

## References

- [CBS97] John Carroll, Ted Briscoe and Antonio Sanfilippo. Parser Evaluation: a Survey and a New Proposal. In *Proceedings of the 1st International Conference on Language Resources and Evaluation*, Granada, Spain. 447-454. <ftp://ftp.cogs.susx.ac.uk/pub/users/johnca/lre98-final.ps>
- [HABFG91] Philip Harrison, Steven Abney, Ezra Black, Dan Flickinger, Claudia Gdaniec, Ralph Grishman, Donald Hindle, Robert Ingria, Mitch Marcus, Beatrice Santorini and Tomek Strzalkowski. Evaluating Syntax Performance of Parser/Grammars of English. *Language Processing Systems Evaluation Workshop, Technical Report RL-TR-91-36, Rome Laboratory, Air Force Systems, Command, Griffis Air Force Base, NY 13441-5700, 1991.*

[Ner98] John Nerbonne. *TMR Network: Learning Computational Grammars project home page*. <http://odur.let.rug.nl/~nerbonne/tmr/>

[RM95] Lance A. Ramshaw and Mitchell P. Marcus. Text Chunking Using Transformation-Based Learning. *Proceedings of the Third ACL Workshop on Very Large Corpora*, 1995. <ftp://ftp.cis.upenn.edu/pub/chunker/wvlcbook.ps.gz>