

1996 IRCS TECHNICAL REPORT ABSTRACTS

A Semantics of Contrast and Information Structure for Specifying Intonation in Spoken Language Generation (Ph.D. Dissertation)
Scott Allan Prevost
IRCS-96-01

In this dissertation I present a model for the determination of intonation contours from context and provide two implemented systems which apply this theory to the problem of generating spoken language with appropriate intonation from high-level semantic representations. The theory and implementations presented here are based on an information structure framework that mediates between intonation and discourse, and encodes the proper level of semantic information to account for both contextually-bound accentuation patterns and intonational phrasing. The structural similarities among these linguistic levels of representation are the basis for selecting Combinatory Categorical Grammar (CCG, Steedman 1985,1990) as the model for spoken language production. This model licenses congruent syntactic, prosodic and information structural constituents and consequently represents a simplification over models of prosody developed in syntactically more traditional frameworks.

The “previous mention” heuristic, which has been widely used as a model for determining intonation contours, is shown to be inadequate for handling a broad range of examples involving semantic contrasts, which require pitch accents to be allocated based on their ability to discriminate among available entities in the discourse model. To address this problem, I introduce a model that determines accentual patterns based on sets of alternative entities in the knowledge base. The algorithms for building the information structural representations that encode the semantics of intonation supply the foundation for two computational implementations. These implementations demonstrate how the theoretical model applies to the problem of producing contextually-appropriate spoken output in a natural language generation framework and provide a platform for incrementally testing and refining the underlying theory.

***k*-Universal Finite Graphs**
Eric Rosen
Saharon Shelah
Scott Weinstein
IRCS-96-02

This paper investigates the class of k -universal finite graphs, a local analog of the class of universal graphs, which arises naturally in the study of finite variable logics. The main results of the paper, which are due to Shelah, establish that the class of k -universal graphs is not definable by an infinite disjunction of first-order existential sentences with a finite number of variables and that there exist k -universal graphs with no k -extendible induced subgraphs.

Critiquing: Effective Decision Support in Time-Critical Domains
Abigail S. Gertner
IRCS-96-03

The TraumAID system is a tool for assisting physicians during the initial definitive management phase of patients with severe injuries. Originally, TraumAID was conceived as a rule-based expert system combined with a planner. After this architecture had been implemented and evaluated, we began to face the issue of how TraumAID could communicate its plans to physicians in order to influence their behavior and have a positive effect on patient outcome. It was hypothesized that a critiquing approach, in which the system is told what actions the user intends to carry out and produces a critique in response to those intentions, might be appropriate.

To meet the needs of physicians engaged in managing trauma cases, critiques must be updated and made available rapidly. They must be clear and succinct, containing only relevant information while still including enough justification to make them convincing. To address the need for these features in the system, I have developed a critiquing architecture consisting of three components: incremental plan recognition, plan evaluation, and critique generation. I have implemented this architecture in TraumaTIQ, a critiquing interface for the TraumAID system. Comparison of TraumaTIQ's comments on 97 actual management plans with comments made by three local trauma experts showed that TraumaTIQ produced 48.3% of comments made by only one judge, and 70.27% of comments made by two or more judges. This relationship between inter-judge agreement and judge-system agreement is statistically significant. In addition, regression analysis shows that TraumaTIQ's plan evaluator is a significant predictor of the judges' overall case ratings.

This approach to communicating with physicians in time-critical domains has the advantage that it is user-focused, minimally intrusive, and quick to respond to potential errors even on the basis of partial information. Unlike previously developed reminder and alert systems, TraumaTIQ evaluates the physician's proposed plan and attempts to intervene before problems occur. And unlike previous critiquing systems, it is able to provide ongoing decision support during the planning and delivery of care. In the context of time-critical patient management it is, therefore, a more appropriate form of interaction.

Mathematical Approaches to Comparative Linguistics
Tandy Warnow
IRCS-96-04

The inference of the evolutionary history of a set of languages is a complex problem. Although some languages are known to be related through descent from common ancestral languages, for other languages determining whether such a relationship holds is itself a difficult problem. In this paper we report on

new methods, developed by linguists Johanna Nichols (Berkeley), Donald Ringe (Penn), and Ann Taylor (Penn), and computer scientist Tandy Warnow (Penn), for answering some of the most difficult questions in this domain. These methods and the results of the analysis based upon these methods were presented in November 1995 at the Symposium on the Frontiers of Science of the National Academy of Science.

The Workings of the Intellect: Mind and Psychology.

Gary Hatfield

IRCS-96-05

This paper examines the importance of the theory of intellectual cognition in the development of early modern philosophy. It compares three conceptions of the intellect, held respectively by some scholastic Aristotelians, Descartes, and Locke. Examination of these three cases provides an opportunity to locate early modern discussions of the cognitive faculties in relation to recent understandings of psychology, epistemology, logic, mind, and their relations. The early modern discussions are not easily fit into the modern categories of epistemology and psychology. Reflection on this fact may help us to delimit more precisely and to see some problems in recent concepts of naturalism in relation to philosophy and psychology.

A Decidable Predicate Logic of Knowledge

Giorgi Japaridze

IRCS-96-06

Defining Cognitive Science at IRCS: Proceedings from the April 1995 Postdoc Workshop

Edited by Anne Vainikka

IRCS-96-07

Language Contact and Language Change in the Faetar Speech Community

Naomi Nagy

IRCS-96-08

This dissertation proposes a standardized approach to the analysis of contact-induced language change and applies the model to an examination of Faetar, a language which has developed in a situation of contact between Italian and Francoprovençal. The model incorporates factors which have been shown to be significant in determining contact-induced change and variation and indicates how these factors can be analyzed at the individual speaker level. Once a uniform approach to the analysis is taken, progress toward a model of contact-induced language change will be more rapid. This approach permits testing of claims regarding the relationship between social context and types of language change.

The language examined is Faetar, an unwritten and virtually unstudied Francoprovençal dialect which has been spoken in a village in southern Italy for the past 600 years. Because the

language has been in contact with Italian as well as the colloquial dialects of the area, it has undergone many changes and is no longer mutually intelligible with Francoprovençal. Unlike many dialects spoken in isolated communities, Faetar does not appear to be dying as the language is held in high regard by its speakers. They recognize the prestige of this marker of their distinctness from other southern Italians. The question arises, however, of just how distinct the language is. In its centuries of contact with Italian, Faetar has changed in many ways. These contact-induced changes, and methods of analyzing them, are the focus of this dissertation.

Four phenomena that show the effects of Italian and Apulian dialects on Faetar are examined. Quantitative analysis of each pattern within a variationist framework, using both elicited and naturally occurring speech data, is presented. The variables examined are the appearance of word-medial and word-initial geminates, post-tonic deletion, and lexical choice. Degree of change at the lexical, phonological, and morphological levels is compared.

A detailed description of these phenomena augments the scarce data available for this uncodified language variety, contributing to the reconstruction of Francoprovençal as it was spoken some 600 years ago, as well as aiding the preservation of Faetar.

Acquisition of Variable Rules: (-t,d) Deletion and (ing) Production

Julia Lee Roberts

IRCS-96-09

There have been many studies over the past few decades documenting the existence of variable rules in adult language. It is only recently, however, that the acquisition of these rules has been the focus of research, and that event has opened the door for questions about the interaction of the learning of categorical rules and that of variable rules. Specifically, questions have arisen as to whether these rules might not be construed as either a performance factor and/or a reflection of universal constraints on language.

The present study examines the acquisition of (-t,d) deletion and (ing) production in 3- and 4-year-old children in order to ascertain their degree of mastery of phonological, grammatical, and social constraints. Seventeen children were tape recorded during play interview sessions in their South Philadelphia day care center. Six to thirteen sessions per child over a three month period were required to obtain sufficient data for analysis. In addition, eight of their parents were interviewed in their homes for purposes of comparison.

Results of the study revealed that children as young as three had, for the most part, mastered the process of variation of (ing) and the phonological constraints on (-t,d) deletion, and they were well into the process of acquiring the grammatical constraints on (-t,d) deletion. Their learning of a dialect

specific phonological constraint demonstrated that their mastery of this variable rule was not a reflection of universal constraints. Further, their independent analysis of semi-weak verbs made it clear that they were not simply copying frequencies of their parents' forms but learning an abstract rule. The children's acquisition of the extralinguistic constraints on these rules lagged behind that of the linguistic factors. Of particular interest to the issue of gender differences in language was the girls' surprising tendency to delete (-t,d) more often than the boys, demonstrating that they had not yet learned linguistic conservatism in instances of stable variation and arguing against a biological basis for sex-based sociolinguistic differences.

What Does a Grammar Formalism Say About a Language?

James Rogers
IRCS-96-10

Over the last ten or fifteen years there has been a shift in generative linguistics away from formalisms based on a procedural interpretation of grammars towards constraint-based formalisms---formalisms that define languages by specifying a set of constraints that characterize the set of well-formed structures analyzing the strings in the language. A natural extension of this trend is to define this set of structures model-theoretically---to define it as the set of mathematical structures that satisfy some set of logical axioms. This approach raises a number of questions about the nature of linguistic theories and the role of grammar formalisms in expressing them.

We argue here that the crux of what theories of syntax have to say about language lies in the abstract properties of the sets of structures they license. This is the level that is most directly connected to the empirical basis of these theories and it is the level at which it is possible to make meaningful comparisons between the approaches. From this point of view, grammar formalisms (or formal frameworks) are primarily means of presenting these properties. Many of the apparent distinctions between formalisms, then, may well be artifacts of their presentation rather than substantive distinctions between the properties of the structures they license. The model-theoretic approach offers a way in which to abstract away from the idiosyncrasies of these presentations.

Having said that, we must distinguish between the class of sets of structures licensed by a linguistic theory and the set of structures licensed by a specific instance of the theory---by a grammar expressing that theory. Theories of syntax are not simply accounts of the structure of individual languages in isolation, but rather include assertions about the organization of the structure of human languages in general. These universal aspects of the theories present two challenges for the model-theoretic approach. First, they frequently are not properties of individual structures, but are rather properties of sets of structures. Thus, in capturing these model-theoretically

one is not defining sets of structures but is rather defining classes of sets of structures; these are not first-order properties. Secondly, the universal aspects of linguistic theories are frequently not explicit, but are consequences of the nature of the formalism that embodies the theory. In capturing these one must develop an explicit axiomatic treatment of the formalism. This is both a challenge and a powerful benefit of the approach. Such re-interpretations tend to raise a variety of issues that are often overlooked in the original formalization.

In this report we examine these issues within the context of a model-theoretic reinterpretation of Generalized Phrase-Structure Grammar. While there is little current active research on GPSG, it provides an ideal laboratory for exploring these issues. First, the formalism of GPSG is expressly intended to embody a great deal of the accompanying linguistic theory. Thus it provides a variety of opportunities for examining principles expressed as restrictions on the formalism from a model-theoretic point of view. At the same time, the fact that these restrictions embody universal grammar principles provides us with a variety of opportunities to explore the way in which the linguistic theory expressed by a grammar can transcend the mathematical theory of the structures it licenses. Finally, GPSG, although defined declaratively, is a formalism with restricted generative capacity_a characteristic more typical of the earlier procedural formalisms. As such, one component of the theory it embodies is a claim about the language-theoretic complexity of natural languages. Such claims are difficult to establish for any of the constraint-based approaches to grammar. We can show, however, that the class of sets of trees that are definable within the logical language we employ in reformalizing GPSG is nearly exactly the class of sets of trees definable within the basic GPSG formalism. Thus we are able to capture the language-theoretic consequences of GPSG's restricted formalism by employing a restricted logical language.

An Empirical Comparison of Probability Models for Dependency Grammar

Jason Eisner
IRCS-96-11

This technical report is an appendix to Eisner (1996): it gives superior experimental results that were reported only in the talk version of that paper, with details of how the results were obtained. Eisner (1996) trained three probability models on a small set of about 4,000 conjunction-free, dependency-grammar parses derived from the Wall Street Journal section of the Penn Treebank, and then evaluated the models on a held-out test set, using a novel $O(n^3)$ parsing algorithm.

The present paper describes some details of the experiments and repeats them with a larger training set of 25,000 sentences. As reported at the talk, the more extensive training yields greatly improved performance, cutting in half the error rate of Eisner (1996). Nearly half the sentences are parsed

with no misattachments; two-thirds of sentences are parsed with at most one misattachment.

Of the models described in the original paper, the best score is obtained with the generative ``model C,’’ which attaches 87-88% of all words to the correct parent. However, better models are also explored, in particular, two simple variants on the comprehension ``model B.’’ The better of these has an attachment accuracy of 90%, and (unlike model C) tags words more accurately than the comparable trigram tagger.

If tags are roughly known in advance, search error is all but eliminated and the new model attains an attachment accuracy of 93%. We find that the parser of Collins (1996), when combined with a highly-trained tagger, also achieves 93% when trained and tested on the same sentences. We briefly discuss the similarities and differences between Collins’s model and ours, pointing out the strengths of each and noting that these strengths could be combined for either dependency parsing or phrase-structure parsing.

The Myth of Empty Subjects: A cross-linguistic perspective with special emphasis on the acquisition of the ‘mixed’ system of Hebrew

Yonata Levy
Anne Vainikka
IRCS-96-12

Studies of the early stages of language acquisition universally show children’s preponderance to omit NP subjects over and above omission of other NP constituents such as direct objects. In a recent paper, Vainikka & Levy (1995) propose an account of null NP subjects which crucially depends on the assumption that the syntactic position of inflectional features varies cross-linguistically and even within a language, such that the possibility exists for these features to occur in the subject position. Consequently, fully inflected languages which license the omission of NP subjects are NOT null-subject languages. The proposed model accounts not only for languages of the English and the Italian type, but it likewise, predicts the mixed pattern of omission that is found in adult Hebrew and Finnish.

With respect to acquisition, this model predicts that knowledge of linguistic person distinctions will enable the child to consider the first and most important clue with respect to the grammar of NP subject omission, namely, the possibility (or the impossibility) of NP subject omission in the 3rd person. Furthermore, control of the forms of personal pronouns, at least in the singular, as well as productive knowledge of the inflectional paradigms, at least in the singular, are required for mastering the adult mixed pattern of NP subject omission in languages such as Hebrew or Finnish. Discovering the phonological relatedness of pronouns and inflectional affixes serves as a central clue to the child in the acquisition of these mixed systems. Longitudinal data from two Hebrew speaking children supporting such an

acquisitional course is presented. Cross-linguistic studies have shown that following these syntactic developments, in which subjects become obligatory and a language-specific pattern of omission has been acquired, there remains a small residue of unacceptable NP subject omissions. It is proposed that this residue is a consequence of children’s over-generalizations of the pragmatic contexts in which omissions are accepted in the adult language. Data from English and Hebrew is presented in support of the proposed analysis.

Null vs. Overt Subjects in Turkish Discourse: A Centering Approach (Ph.D. Dissertation)

Umit Deniz Turan
IRCS-96-13

The purpose of this study is to explore an aspect of discourse coherence which involves anaphoric relations between utterances with special emphasis on subjects in Turkish. Based on an analysis of published narratives, three complementary and interrelated questions are addressed concerning discourse anaphora:

1. Which expressions are available for subsequent definite reference?
2. What factors determine the most salient entity in Turkish among a set of potential antecedents for subsequent definite reference?
3. What are the functions of a particular referential expression (null vs. overt pronouns vs. full NPs), depending on appropriate discourse conditions?

An exploration regarding question 1 indicates that, while some NPs evoke discourse entities, other NPs do not. These two types of NPs represent referential and nonreferential expressions and they can function as antecedents for definite and indefinite nonspecific anaphora, respectively. The distinction between null and overt pronouns in Turkish is that only the former can be in an anaphoric relationship with a nonreferential antecedent. Overt pronouns, on the other hand, are sensitive to referent identity: they must have the same referent with their antecedents. In other words, overt pronouns are strictly coreferential, while null pronouns are not constrained in this way.

The rest of the study investigates answers to questions 2 and 3 in instances where null and definite subjects alternate as definite anaphors. Centering Theory provides a cognitively plausible and computationally tractable framework for such an analysis with its precise rules which rely on linguistic knowledge constraining inferencing. As formulated in Centering Theory, each utterance contains a set of potential antecedents for reference in the subsequent utterance, i.e. a set of forward-looking centers (Cfs), that are ranked on the basis of their salience. The most salience entity in the Cf-list, the

preferred center (Cp), is the entity that is predicted to become the backward-looking center in the subsequent utterance. The singleton backward-looking center (Cb) is taken to be the topic of the current utterance, i.e. the entity at the center of attention.

Centering transitions, which model the dynamic attentional state in a discourse segment, are obtained by analyzing each adjacent pair of utterances. The functions of referential expressions in subject position are determined on the basis of Centering transitions. The results show that Turkish subject types pattern neatly and categorically when these transitions are taken into account.

A brief discussion of language-specific and universal aspects of discourse anaphora is also included in the study.

Complexity and the Induction of Tree-Adjoining Grammars

Robin Clark
IRCS-96-14

In this paper, I will develop the formal foundations of a theory of complexity that underlies theory of grammatical induction. The initial concern will be the learning theoretic foundations of linguistic locality. That is, I will develop a theory that will place bounds on the amount a learner can draw from an input text. These bounds will limit the amount of variation that could potentially be encoded within a parameter space. A fully developed form of the theory will place a tangible upper limit on what the learner can induce from the input text. The formal theory developed establishes a relationship between the complexity of descriptions and their likelihood; that is, the more complex a structure is, the less likely it is to occur. I will use this result to develop a theory of linguistic complexity. I will rely on this relationship to show that the results developed in the first part of the paper for the parameter setting model also hold for the inductive theory. The final sections of the paper turn to the formal specification of the learning model and a description of the linguistic theory that supports it. This section also describes a pair of heuristic constraints on the learner's search for viable hypotheses. In general, the learner faces a computationally intractable problem in that there are exponentially many grammatical hypotheses for any input text. These constraints, the Adjunction Constraint and the Substitution Constraint, greatly reduce the number of hypotheses that the learner must consider. Furthermore, metrics on the complexity of the learner's descriptions guarantee that the hypothesis space can be tractably searched for the adult grammar.

The Prosodic Structure of Latvian **(Ph.D. Dissertation)**

A. Krisjanis Karins
IRCS-96-15

This dissertation investigates the rhythmic and melodic

structure of standard Latvian, which has a system of syllable intonations independent of the metrical structure of the language. Latvian can thus be classified as a semi-tonal language which is between a pitch-accent language and a purely tonal language. The investigation of the prosodic structure is based upon empirical data gathered and analyzed using techniques of experimental phonetics, and builds upon earlier linguistic work on Latvian conducted primarily within the tradition of Latvian philology. The analysis of the data is conducted within the generative frameworks of Optimality Theory and autosegmental phonology.

The phonology of Latvian distinguishes "long" syllables from "heavy" ones. Whereas metrically long syllables can influence the duration of voiceless obstruents, only metrically heavy syllables can be associated with lexical tones. This division of syllable types justifies a two-layer moraic analysis of the language.

The metrical structure of the language ideally builds two feet or one colon per word, with the addition that the main word stress is associated with a H tone. Of the three syllable intonations-level, falling, and broken-only the falling and broken are lexically specified for tone. The level intonation is lexically unspecified, and its tonal contour is dependent upon the presence or absence of a stress-induced H tone. Thus, although the metrical and tonal systems are independent in the language (every heavy syllable has a syllable intonation regardless of stress), they nevertheless interact insofar as the metrical H tone influences the tonal contours of all three syllable intonations.

The Question of Root Infinitives in Early Child Greek **Spyridoula Varlokosta, Anne Vainikka, & Bernard Rohrbacher** **(Brown University)** **IRCS-96-16**

It is well known that children acquiring Germanic and Romance languages go through an early stage at which they produce declarative sentences with a Root Infinitive (cf. (1)) which would be ungrammatical in the adult language (Stern & Stern 1928, Weverink 1989, Pierce 1992, Wexler 1994). For languages outside of these language families, little work has been done on such Root Infinitives. In the present paper we investigate the status of Root Infinitives in Modern Greek, a language which lacks an infinitive form altogether.

Transforming Databases with Recursive Data Structures **(Ph.D. Dissertation)** **Anthony Kosky** **IRCS-96-17**

This thesis examines the problems of performing structural transformations on databases involving complex data-structures and object-identities, and proposes an approach to specifying and implementing such transformations. We start by looking at various applications of such database

transformations, and at some of the more significant work in these areas. In particular we will look at work on transformations in the area of database integration, which has been one of the major motivating areas for this work. We will also look at various notions of correctness that have been proposed for database transformations, and show that the utility of such notions is limited by the dependence of transformations on certain implicit database constraints. We draw attention to the limitations of existing work on transformations, and argue that there is a need for a more general formalism for reasoning about database transformations and constraints. We will also argue that, in order to ensure that database transformations are well-defined and meaningful, it is necessary to understand the information capacity of the data-models being transformed. To this end we give a thorough analysis of the information capacity of data-models supporting object identity, and will show that this is dependent on the operations supported by a query language for comparing object identities.

We introduce a declarative language, WOL, based on Horn-clause logic, for specifying database transformations and constraints. We also propose a method of implementing transformations specified in this language, by manipulating their clauses into a normal form which can then be translated into an underlying database programming language.

Finally we will present a number of optimizations and techniques necessary in order to build a practical implementation based on these proposals, and will discuss the results of some of the trials that were carried out using a prototype of such a system.

Graphical Communicating Shared Resources: A Language for Specifying, Refining, and Analyzing Real-Time Systems (Ph.D. Dissertation)
Hanène Ben Abdallah
IRCS-96-18

The Communicating Shared Resources (CSR) paradigm is an ongoing project at the University of Pennsylvania to build a framework for the development of real-time systems. This project has been motivated by a demand for a rigorous framework in which various design alternatives for a real-time system can be formally specified and rigorously analyzed and tested before implementation. This is an effort to reduce the potentially high cost associated with incorrect operation of real-time systems which are often embedded in safety-critical applications.

The work presented in this thesis is a first step towards incorporating software engineering practices into the CSR paradigm. This is achieved, on one hand, by developing a formal, graphical CSR formalism, the Graphical Communicating Shared Resources (GCSR); the GCSR language adopts the intuitive concepts of nodes and edges in state diagrams, an informal specification language that is

popular within the software engineering community. In addition, defining a refinement theory for GCSR allows the development of real-time systems within this formalism in a top-down and modular fashion, also a popular design methodology within the software engineering community. The GCSR language adopts a syntax that allows a modular and hierarchical, thus, scalable description of a real-time system. It supports notions of communication through events, interrupt, concurrency, and time to describe the functional and temporal requirements of a real-time system. In addition, GCSR allows the explicit representation of resources and priorities to resolve resource contention, in such a way that produces easy to understand and modify specifications. The semantics of GCSR is defined operationally either through a direct translation of a GCSR description to a labeled transition system, or indirectly through a sound translation to the Algebra of Communicating Shared Resources (ACSR) [LBGG94], a timed process algebra that also has an operational semantics. The GCSR-ACSR correspondence makes GCSR benefit from process algebraic analysis techniques such as equivalence checking, state space exploration, testing as well as simulation. In addition, the tight correspondence between GCSR and ACSR makes it possible to use the graphical and textual notations interchangeably and to have a sound theory for graphical transformation operations, e.g., to minimize the number of edges and nodes in a GCSR specification without affecting the behavioral description.

To support the top-down and modular development of a real-time specification in GCSR, we have augmented ACSR and thus GCSR with a refinement theory. The refinement theory allows relabeling of events, addition of implementation events, and substitution of a time and resource-consuming action with a process that may use fewer or more resources than the refined action. Consistency between an abstract specification and a refined specification is defined in terms of an ordering relation over traces that is extended to sets of traces according to the Hoare ordering or Egli-Milner ordering. The trace ordering relation relates traces that share timing properties such as equal duration and preservation of timed occurrences of communication events of the abstract specification. To facilitate the practical use of the refinement theory, we have characterized the extended trace ordering relations by a set of transformation rules that syntactically derive a refined process from an abstract one. The transformation rules define basic graphical operations that represent GCSR refinements.

To experiment with the GCSR language and its refinement theory, we have developed a tool set that allows the specification, refinement, and analysis of real-time systems modeled in GCSR. We report our evaluation in the case of the Production Cell case study [LL95].

Learning First Order Quantifier Denotations: An Essay in Semantic Learnability

Robin Clark
IRCS-96-19

This paper addresses the problem of how a learner would associate a denotation with a determiner on the basis of the pairing of a simple sentence with some perceptual input. Following work by van Benthem (1985), first order quantification is simulated with (a subclass of) finite state automata. Results from Angluin (1988) are used to demonstrate the learnability of this class.

Variable Zero-Marking of (o) in Tokyo Japanese (Ph.D. Dissertation)

Kenjiro Matsuda
IRCS-96-20

The variation between the presence and the absence of the object case marker -o (the zero-marking of (o)) in Tokyo Japanese has been approached from a number of different linguistic traditions, including generative grammar, functionalism, sociolinguistics and corpus linguistics, and a number of hypotheses have been advanced concerning the constraints governing this phenomenon of natural speech. Nevertheless, no comprehensive studies on the phenomenon have yet been done with natural speech as the primary source of data. In this variationist thesis, I select factors from major proposals in recent literature, and demonstrate through a multivariate analysis that the variation can be accounted for by (1) the form of the object NP, (2) word order and (3) a factor encoding the interaction between residence area (uptown and downtown) and speech style. The object NP form factor is shown to obey the following hierarchy: win-pronoun > NP > non-w in-pronoun > clausal NP (in order of decreasing likelihood of zero-marking), reflecting a crosslinguistic tendency for pronouns to be case-marked. Adjacency between the verb and the object NP is the most crucial aspect of the word order factor. Furthermore, a close examination of the Extended Functional Hypothesis reveals that the variation is not driven by any functional motivation, such as the need to preserve meaning, but is rather the result of a mechanical process whereby a zero form is followed by another zero, a conclusion that supports results from variable processes in other languages (Poplack 1979, Scherre and Naro 1991). On the social dimension, women predominate in the use of the zero form in both residential areas, but uptown speakers of both sexes exhibit sharper style shifting than their downtown counterparts, a point that was overlooked in previous work on the variable (Shibamoto 1985). This social differentiation is explained in this thesis as a reflection of the historical background of each residential area. Finally, I note that it cannot at present be determined whether this instance of synchronic variation is a case of stable variation or of (slow) change in progress.

On the Phonology and Morphology of Andalusian Spanish: A Phenomenon of Vowel Raising

Fabiola Varela-Garcia
IRCS-96-21

The variable effects of final /-s/ in the nominal and the verbal morphology of Romance languages have been profusely addressed in the linguistic literature. Phonetically, lengthening of final /-s/ results in greater articulatory efficiency in a number of Romance languages such as French, Italian and Spanish. Vowel lengthening, lowering and raising have been postulated as among the most frequent compensatory mechanisms after morpheme loss.

Intermediate stages, such as aspiration or deletion of /-s/ and vowel lengthening have been attested in French, Occitan and Spanish. Vocalization of /-s/ in monosyllabic words has been posited as the origin of the vocalic alternation that presently denotes plurality in Italian. However, vowel lowering and raising have only been reported in Spanish, although raising has also been postulated for Romanian, Occitan and Italian (Seklaoui 1989).

Fieldwork conducted in Estepa (Seville), Spain, provides the data for the present study. This dissertation is the first quantitative report within a Labovian variationist approach to document the sociolinguistic constraints affecting vowel raising in Andalusian Spanish. The analysis of linguistic factors is construed so as to reveal any possible compensatory effect that raising of /a/ may have following final /-s/ deletion in such noun phrases as *las habichuelas rojas*, 'the red beans.' In the same fashion, a functional effect is hypothesized for vowel raising in the verbal morphology. Effectively, the absence of a disambiguating subject pronoun and the deletion of /-s/ would leave /a/ raising as the only plural marker in instances like *juegas mucho* 'you -singular-play a lot'.

A multivariate analysis selects syllable count as the most important factor constraining vowel raising. The monosyllabic determiner *las* remains almost categorically unraised, whereas nouns, adjectives and the disyllabic pronoun *ellas* have the highest probability of vowel raising. The probabilistic analysis also selects final /-s/ as a significant factor for /a/ raising in Andalusian Spanish. However, although it points out the correlation between /a/ raising and /-s/ deletion, the analysis also attests the change of /a/ to [e] in the presence of an aspirated segment, as in [lah amigeh míeh] *las amigas mías*, 'my -female-friends'. Moreover, a serial effect reported for Puerto Rican Spanish (Poplack 1979) and Portuguese (Scherre & Naro, 1991, 1992, 1995) is also attested in this variety of Spanish, not only for /-s/ lenition but also for vowel raising. Analysis at the syntagmatic level reveals the highest percentages of vowel raising in the case of a previous raised vowel within the noun phrase. In addition to this, statistical analysis depicts a functional interval for this vocalic change. In the absence of plural /-s/, the greatest probability for vowel raising occurs in the adjectives, while the least likely to raise

is the determiner *las*. Overall, this analysis of Andalusian Spanish reveals a structure that keeps feminine plural [eø] to the right of the head, and feminine plural [ah] to the left.

The probabilistic analysis also includes an interactive category of social factors comprised by social class, gender, age and education. It reveals that vowel raising is primarily a working class phenomenon which has experienced social leveling over the last forty years. At present, the practice of vowel raising is also conditioned by the choice Estepenians's make between an urban lifestyle associated with the standard [a] pronunciation and the traditional life of Estepa for which [e] is the vernacular norm. Although /a/raising prevails among women, at present the [e] pronunciation denotes an unprecedentedly symbolic value of local identity for young males.

A Geometric Approach to Mapping Bitext

Correspondence

I. Dan Melamed

IRCS-96-22

The first step in most corpus-based multilingual NLP work is to construct a detailed map of the correspondence between a text and its translation. Several automatic methods for this task have been proposed in recent years. Yet even the best of these methods can err by several typeset pages. The Smooth Injective Map Recognizer (SIMR) is a new bitext mapping algorithm. SIMR's errors are smaller than those of the previous front-runner by more than a factor of 4. Its robustness has enabled new commercial-quality applications. The greedy nature of the algorithm makes it independent of memory resources. Unlike other bitext mapping algorithms, SIMR allows crossing correspondences to account for word order differences. Its output can be converted quickly and easily into a sentence alignment. SIMR's output has been used to align over 200 megabytes of the Canadian Hansards for publication by the Linguistic Data Consortium.

Automatic Detection of Omissions in Translations

I. Dan Melamed

IRCS-96-23

ADOMIT is an algorithm for Automatic Detection of Omissions in Translations. The algorithm relies solely on geometric analysis of bitext maps and uses no linguistic information. This property allows it to deal equally well with omissions that do not correspond to linguistic units, such as might result from word-processing mishaps. ADOMIT has proven itself by discovering many errors in a hand-constructed gold standard for evaluating bitext mapping algorithms. Quantitative evaluation on simulated omissions showed that, even with today's poor bitext mapping technology, ADOMIT is a valuable quality control tool for translators and translation bureaus.

Syntax and Discourse Factors in Early New High German: Evidence for Verb-Final Word Order (Master's thesis)

Ann Bies

IRCS-96-24

Several recent studies have taken the approach that the word order variation in Early New High German (1300-1600) is indicative of a change in the underlying syntax. In this thesis, I argue that Early New High German (ENHG) was solidly verb-final in structure throughout its history and that the evidence for the existence of structurally INFL-medial sentences in ENHG is not convincing. I find no evidence of competing grammars in ENHG, and it is therefore notable that the loss of NP postposition between ENHG and modern German appears to be an example of language change due to a change in frequency of usage. In addition, NP postposition is shown to have a particular discourse function: forcing the postposed NP to be interpreted with narrow focus.

ENHG differs syntactically from modern German in allowing surface word orders that are ungrammatical or rare in later German. Three such syntactic constructions are investigated - NP postposition, PP postposition and Verb (Projection) Raising. The postposition of clauses and prepositional phrases occurs in both ENHG and modern German. However, in ENHG, NP objects can move to a position after a structurally final verb, a movement that is not possible in modern German. Verb Raising and Verb Projection Raising, in which the relative order of verb forms is reversed from the standard order, are also much more common constructions in ENHG than in modern German. All of these constructions involve surface word orders which are not verb-last, and each occurs at an overall rate of 24-30% of possible cases in my corpus. The variation shows synchronic social, stylistic and discourse effects.

I argue that in the absence of evidence for competing grammars, the observed variation in string verb-last word order in ENHG should be understood as syntactic variation in an underlyingly structurally verb-final language. The decline of these three unrelated syntactic constructions between ENHG and modern German may be due to the imposition of a standard surface verb-last template from above. Such a change from above forces unrelated structures, such as NP focus postposition, Verb (Projection) Raising and PP postposition, to change in the same direction, in this case toward surface verb-last word order.

Pipeline Rendering: Interaction and Realism Through Hardware-Based Multi-Pass Rendering

(Ph.D. Dissertation)

Paul J. Diefenbach

IRCS-96-25

While large investments are made in sophisticated graphics hardware, most realistic rendering is still performed off-line using ray trace or radiosity systems. A coordinated use of

hardware-provided bit planes and rendering pipelines can, however, approximate ray trace quality illumination effects in a user-interactive environment, as well as provide the tools necessary for a user to declutter such a complex scene. A variety of common ray trace and radiosity illumination effects are presented using multi-pass rendering in a pipeline architecture. We provide recursive reflections through the use of secondary viewpoints, and present a method for using a homogeneous 2-D projective image mapping to extend this method for refractive transparent surfaces. This paper then introduces the Dual Z-buffer, or DZ-buffer, an evolutionary hardware extension which, along with current frame-buffer functions such as stencil planes and accumulation buffers, provides the hardware platform to render non-refractive transparent surfaces in a back-to-front or front-to-back order. We extend the traditional use of shadow volumes to provide reflected and refracted shadows as well as specular light reclassification. The shadow and lighting effects are then incorporated into our recursive viewpoint paradigm. Global direct illumination is provided through a shadow blending technique. Hardware surface illumination is fit to a physically-based BRDF to provide a better local direct model, and the framework permits incorporation of a radiosity solution for indirect illumination as well. Additionally, we incorporate material properties including translucency, light scattering, and non-uniform transmittance to provide a general framework for creating realistic renderings. The DZ-buffer also provides decluttering facilities such as transparency and clipping. This permits selective scene viewing through arbitrary view-dependent and non-planar clipping and transparency surfaces in real-time. The combination of these techniques provide for understandable, realistic scene rendering at typical rates 5-50 times that of a comparable ray trace images. In addition, the pixel-parallel nature of these methods leads to exploration of further hardware rendering engine extensions which can exploit this coherence.

Porting SIMR to New Language Pairs

I. Dan Melamed

IRCS-96-26

There are three steps in porting SIMR to new language pairs. The first step is to choose an appropriate matching predicate, and collect any linguistic resources required by that matching predicate. The second step is to implement bitext space axis generation routines that are consistent with the matching predicate. The last step is to re-optimize SIMR's numerical parameters. This document explains each step in detail. It assumes that you have read and understood the paper [1]. It also assumes that you have SIMR and the porting tools properly installed.

A Lexical Theory of Quantification in Ambiguous Query Interpretation (Ph.D. Dissertation)

Jong Cheol Park

IRCS-96-27

Although the connection between natural language syntax and semantics has received serious attention in both linguistics and computational linguistics for the last several decades, it does not appear that it has yet been entirely satisfactorily identified. The present dissertation focuses on quantifier scope ambiguity in an attempt to identify such a connection. We show that there are some readings that are incorrectly allowed by the theories and that other readings that are available are allowed for the wrong reason.

First, we distinguish referential NP interpretations from quantificational NP interpretations. Most traditional theories of scope do not, and they are shown to significantly overgenerate readings and/or miss a crucial generalization regarding quantificationally available readings. We present a hypothesis based on the notion of surface constituency to predict quantificationally available readings. The hypothesis is tested on core English constructions, including transitive verbs, dative alternation (ditransitive) verbs, attitude verbs, complex NPs containing prepositional phrases, possessives, and subject or non-subject {em Wh}-relatives (also with pied-piping), and various coordinate structures. We argue that the scopings allowed under the hypothesis are the ones that are available.

We then present a competence theory of quantifier scope, couched in a combinatory categorial grammar framework. The theory defines the connection between syntax and semantics in a precise way, utilizing the dual quantifier representation. We show theoretical predictions on the core English constructions, and verify that the theoretical predictions are consistent with the predictions made by the hypothesis and that there are further reasonable theoretically predicted readings.

Finally, we describe an implementation of the theory in Prolog. The implemented system takes English sentences as ambiguous queries (regarding scope), generates logical forms that are associated with them, and evaluates those logical forms with respect to a predefined database of facts. The system also works as a proof-checker of the theory.

Quantifier Scope, Lexical Semantics, and Surface Structure Constituency

Jong C. Park

IRCS-96-28

We present a novel conjecture concerning the scope ambiguities that arise in sentences including multiple non-referential quantifiers. We claim that many existing theories of the phenomenon fail to correctly limit the set of readings that such sentences engender by failing to distinguish between referential and non-referential quantifiers. Once the distinction is correctly drawn, we show that surface syntax can be made, via an extended notion of surface constituency, to identify the set of available differently-scoped readings for such sentences. We examine various English constructions to show that the scopings predicted by the conjecture are the only ones that are

available to human language understanders. We show how to incorporate this conjecture into a theory of quantifier scope, by couching it in a unification-based Combinatory Categorical Grammar framework and implementing it in SICStus Prolog. Finally, we compare the proposal with related approaches to quantifier scope ambiguity.

**Using Syntactic Information in Document Filtering:
A Comparative Study of Part-of-speech Tagging and
Supertagging**
R. Chandrasekar and B. Srinivas
IRCS-96-29

Any coherent text contains significant latent information, such as syntactic structure and patterns of language use. This information can be exploited to overcome the inadequacies of keyword-based retrieval and make information retrieval more efficient. In this paper, we demonstrate quantitatively how syntactic information is useful in filtering out irrelevant documents. We also compare two different syntactic labelings -- simple Part-of-speech (POS) labeling and Supertag labeling -- and show how the richer (more fine-grained) representation of supertags leads to more efficient and effective document filtering. We have implemented a system which exploits syntactic information in a flexible manner to filter documents. The system has been tested on a large collection of news sentences, and achieves an F-score of 89 for filtering out irrelevant sentences. Its performance and modularity makes it a promising postprocessing addition to any Information Retrieval system.

Automatic Induction of Rules for Text Simplification
R. Chandrasekar and B. Srinivas
IRCS-96-30

Long and complicated sentences pose various problems to many state-of-the-art natural language technologies. We have been exploring methods to automatically transform such sentences as to make them simpler. These methods involve the use of a rule-based system, driven by the syntax of the text in the domain of interest. Hand-crafting rules for every domain is time-consuming and impractical. This paper describes an algorithm and an implementation by which generalized rules for simplification are automatically induced from annotated training material with a novel partial parsing technique which combines constituent structure and dependency information. This algorithm described in the paper employs example-based generalizations on linguistically-motivated structures.

**Gleaning information from the Web: Using Syntax to
Filter out Irrelevant Information**
R. Chandrasekar and B. Srinivas
IRCS-96-31

In this paper, we describe a system called Glean, which is predicated on the idea that any coherent text contains

significant latent information, such as syntactic structure and patterns of language use, which can be used to enhance the performance of Information Retrieval systems. We propose an approach to information retrieval that makes use of syntactic information obtained using a tool called a supertagger. A supertagger is used on a corpus of training material to semi-automatically induce patterns that we call augmented-patterns. We show how these augmented patterns may be used along with a standard Web search engine or an IR system to retrieve information, and to identify relevant information and to filter out irrelevant items. We describe an experiment in the domain of official appointments, where the use of such patterns to filter out irrelevant documents leads to a reduction of the potential space of matches by upwards of 60% .

**Kinematic Control of Human Postures for Task
Simulation (Ph.D. Dissertation)**
Xinmin Zhao
IRCS-96-32

Kinematic control of human postures for task simulation is important in human factor analysis, simulation and training. It is a challenge to control the postures of a synthesized human figure in real-time on today's graphics workstations because the human body is highly articulated. In addition, we need to consider many spatial restrictions imposed on the human body while performing a task.

In this study, we simplify the human posture control problem by decoupling the degrees of freedom (dof) in the human body. Based on several decoupling schemes, we develop an analytical human posture control algorithm. This analytical algorithm has a number of advantages over existing methods. It eliminates the local minima problem, it is efficient enough to control whole human body postures in real-time, and it provides more effective and convenient control over redundant degrees of freedom. The limitation of this algorithm is that it cannot handle over-constrained problems or general constraint functions. To overcome this limitation, we transform the human posture control problem from a 40 variable joint space to a 4 to 9 redundancy parameter space. We then apply nonlinear optimization techniques on the transformed problem. Because the search space is reduced, this new numerical algorithm is more likely to find a solution than existing methods which apply optimization techniques directly in the joint space.

The contributions of this thesis include a decoupling approach for simplifying the human posture control problem, an analytical human posture control algorithm based on this decoupling approach, and a numerical human posture control algorithm in redundancy parameter space. These two new algorithms are more efficient and effective than existing methods, and they also give the user control to select the desired solution. Moreover, the analytical algorithm can control postures of a few 92 dof human figures at 30 Hz.

Model-based Shape and Motion Analysis: Left Ventricle of a Heart (Ph.D. Dissertation)

Jinah Park
IRCS-96-33

The accurate and clinically useful estimation of the shape, motion, and deformation of the left ventricle of a heart (LV) is an important yet open research problem. Recently, computer vision techniques for reconstructing the 3D shape and motion of the LV have been developed. The main drawback of these techniques, however, is that their models are formulated in terms of either too many local parameters that require non-trivial processing to be useful for close to real time diagnosis, or too few parameters to offer an adequate approximation to the LV motion.

To address the problem, we present a new class of volumetric primitives for a compact and accurate LV shape representation in which model parameters are functions. Lagrangian dynamics are employed to convert geometric models into dynamic models that can deform according to the forces manifested in the data points. It is thus possible to make a precise estimation of the deformation of the LV shape (endocardial, epicardial and anywhere in between) with a small number of intuitive parameter functions.

We believe that the proposed technique has a wide range of potential applications. In this thesis, we demonstrate the possibility by applying it to the 3-D LV shape and motion characterization from magnetic tagging data (MRI-SPAMM). We show that the results of our experiments with normal and abnormal heart data enable us to quantitatively verify the physicians' qualitative conception of the left ventricular wall motion.

Continuous Methods for Motion Planning (Ph.D. Dissertation)

Milos Zefran
IRCS-96-34

Motion planning for a robotic system addresses the problem of finding a trajectory and actuator forces that are consistent with a given set of constraints and perform a desired task. In general, the problem is under-determined and admits a large number of solutions. The main claim of this dissertation is that a natural way to resolve the indeterminacy is to define performance of a motion and find a solution with the best performance. The motion planning problem is thus formulated as a variational problem. The proposed approach is continuous in the sense that the motion planning problem is not discretized.

A distinction is made between dynamic and kinematic motion planning. Dynamic motion planning provides the actuator forces as part of the motion plan and requires finding a motion that is consistent with the dynamic equations of the system,

satisfies a given set of equality and inequality constraints, and minimizes a chosen cost functional. In kinematic motion planning, dynamic equations of the system are not taken into account and it is therefore simpler.

For dynamic motion planning, a novel numerical method for solving variational problems is developed. The continuous problem is discretized by finite-element methods and techniques from nonlinear programming are used to solve the resulting finite-dimensional optimization problem. The method is employed to find smooth trajectories and actuator forces for two planar cooperating manipulators holding an object. The computed trajectories are used to model human motions in a two-arm manipulation task. The method is then extended for systems that change the dynamic equations as they move. An example of a simple grasping task illustrates that for such systems variational approach unifies motion planning and task planning.

Kinematic motion planning is formulated as a variational problem on the group of spatial rigid body displacements, $SE(3)$. A Riemannian metric and an affine connection are introduced to define cost functionals that measure smoothness of trajectories. Metrics and connections that are important for kinematic analysis are identified. It is then shown how the group structure of $SE(3)$ can be used to find smooth trajectories that satisfy boundary conditions on positions, orientations, velocities or their derivatives, and have certain invariance properties with respect to the choice of the inertial and body fixed frames.

Computational Methods for Realistic Image Synthesis (Ph.D. Dissertation)

Min-Zhi Shao
IRCS-96-35
MS-CIS-96-36

In this thesis, we investigate the computational methods for both diffuse and general reflections in realistic image synthesis and propose two new approaches: the overrelaxation solution and the Bernstein polynomial solution.

One of the major concerns with the radiosity method is its expensive computing time and memory requirements. In this thesis, we analyze the convergence behavior of the progressive refinement radiosity method and propose two overrelaxation algorithms: the gathering and shooting solution and the positive overshooting solution. We modify the conventional shooting method to make the optimal use of the visibility information computed in each iteration. Based on a concise record of the history of the unshot light energy distribution, a solid convergence speed-up is achieved.

Though a great effort has been made to extend the radiosity method to accommodate general non-diffuse reflection, the current algorithms are still quite limited to simple environment settings. In this thesis, we propose using the piecewise

spherical Bernstein basis functions over a geodesic triangulation to represent the radiance function. The representation is intrinsic to the unit sphere, and can be efficiently stored, evaluated, and subdivided by the numerically stable de Casteljau algorithm. We demonstrate that the computation of other fundamental radiometric quantities such as vector irradiance and reflected radiance can be reduced to the integration of the piecewise spherical Bernstein basis functions. A novel geometric integration algorithm based on adaptive domain subdivision is presented for the Bernstein-Bezier polynomials over a geodesic triangle on the unit sphere.

Quantification and Predication in Mandarin Chinese: A Case Study of Dou (Ph.D. Dissertation)
Shi-Zhe Huang
IRCS-96-36

In the more recent generalized quantifier theory, ‘every’ is defined as a relation between two sets such that the first set is a subset of the second set (Cooper (1987), van Benthem (1986)). We argue in this dissertation that the formal definition of ‘every’ ought to reflect our intuition that this quantifier is always associated with a pairing. For instance, ‘Every student left’ means that for every student there is an event (Davidson (1966), Kroch (1974), Mourelatos (1978), Bach (1986)) such that the student left in that event.

We propose that the formal translation of EVERY be augmented by relating its two arguments via a skolem function. A skolem function links two variables by making the choice of a value for one variable depend on the choice of a value for the other. This definition of EVERY, after which ‘every’ and its Chinese counterpart ‘mei’ can be modeled, can help us explain the co-occurrence pattern between ‘mei’ and the adverb ‘dou’.

It was observed in S.-Z. Huang (1995a) that ‘mei’ requires either ‘dou’, or an indefinite phrase, or a reflexive in its scope. Under the skolemized definition of EVERY, this is explainable: The skolem function needs a variable in the scope of EVERY. We stipulate that only morphologically/lexically licensed variables are available for quantification (of this kind). ‘Dou’ occurs with ‘mei’ because ‘dou’ can license the event variable for skolemization. This function of ‘dou’ is performed by the tense operator in English, while Chinese, lacking tense, resorts to ‘dou’.

‘Dou’, we will argue, is a sum operator on the event variable. Thus, ‘dou VPs’ always assert plural events, which predicts that the distribution of ‘dou’ may or may not involve universal quantification. Among other things, our account explains scope ambiguity in Chinese, the optionality of ‘dou’, and the interchangeability, in a number of contexts, between ‘dou’ and conjunction/additive words for VPs such as ‘ye’ “also, and”, ‘you’ “also, again”, and ‘hai’ “also, still”.

Parsing MELL Proof Nets
Stefano Guerrini
Andrea Masini (Università di Pisa)
IRCS-96-37

We propose a new formulation for full (weakening and constants included) multiplicative and exponential (MELL) proof nets, allowing a complete set of rewriting rules to parse them. The recognizing grammar defined by such a rewriting system (confluent and strong normalizing on the new proof nets) gives a correctness criterion that we show equivalent to the Danos-Regnier one.