Proceedings of the

First Workshop on

Psycho-computational Models of Human Language Acquisition

Held in cooperation with COLING-2004

28-29 August 2004
Geneva, Switzerland

Introduction

Every day, we use language so effortlessly that we often overlook its complexity. The fact that language *is* complex is indisputable. Indeed, even after decades of scrutiny, highly-trained adult scientists cannot agree on a definitive analysis of the underlying mechanism that ultimately determines how our sounds, words, and sentences go together – but such an effortless task for a child! Children as young as one-and-a-half-years-old (and younger) continually exploit much of language's underpinnings while going about the business of making sense of the linguistic environment that surrounds them. By the time a child reaches kindergarten, he or she has almost full mastery of an elaborate structure that eludes adequate scientific description. How children accomplish this – how they come to acquire 'knowledge' of language's essential organization – is one of the most fundamental, beguiling, and surprisingly open questions of modern science.

This workshop brings together researchers whose (at least one) line of investigation is to computationally model the acquisition process and ascertain substantive interrelationships between a model and linguistic and psycholinguistic theory. Progress in this agenda not only directly informs developmental psycholinguistic and linguistic research, but in my opinion, will also have the long term benefit of informing applied computational linguistics in areas that involve the automated acquisition of knowledge from a human or human-computer linguistic environment.

The level of sophistication and breadth of applied computational linguistics techniques has skyrocketed in the past two decades. There is now a battery of computational formalisms and statistical methods to 'choose from,' all which have yielded remarkable success in many applied domains that involve the computer learning of natural language (e.g. speech recognition, web technologies, corpus analysis, etc). These achievements have dramatically spurred even more research and funding to the point where the evolution of the science of computational linguistics can be seen as quickly outpacing that of psycholinguistics.

However, there are signs that the computational linguistics community has been progressively more aware that language technologies might benefit by incorporating learning strategies employed by humans. Although research involving the psycho-computational modeling of human language acquisition has been long active in the areas of psycholinguistics, cognitive science and formal learning theory, it has, arguably, only recently become a growing part of the computational linguistics agenda. This is evidenced by the occasional special session at an ACL meeting (e.g., ACL-1999 – Thematic Session on Computational Psycholinguistics), current workshops at both COLING-2004 (this workshop) and ACL-2004 (Incremental Parsing: Bringing Engineering and Cognition Together), and regular invitations to developmental psycholinguists to deliver plenary addresses at recent ACL meetings. This cross-discipline attentiveness is clearly very healthy and might well help reduce the possibility that applied research will run into a *psycho-computational bottleneck* – when state-of-the-art computational methods cannot be improved further in the development of user-transparent computer-human language applications – by incorporating theoretical advances in computational psycholinguistics into computational language learning technologies.

This workshop brings together a wide range of computational psycholinguistics research that is involved with the study of language acquisition: 34% of author contributions come from researchers holding positions in computer science or related departments, 33% from linguistics departments, 30% from psychology or cognitive science departments, and 3% from other departments.[1] The articles present investigations involving a broad diversity of formalisms, learning strategies, modeling techniques and linguistic phenomena. Linguistic footings range from (variations on): Universal Grammar, constructionist frameworks, and categorial grammar, to novel formulations of structural representation, to 'none.' Learning strategies include: distributional and corpus techniques, connectionist implementations, cue-based learning, and hybrid models that apply several strategies. Phenomena that are modeled include: the acquisition of semantics, linguistic (principles and) parameter setting, lexical subcategorization, child language production, atypical acquisition, phonological acquisition and morphological acquisition. Several papers involve cross-linguistic research and/or use actual child-directed speech (from corpora).

[1] An "author contribution" is calculated as 1 / the number of authors on a paper.

Notably, most papers (not all) address acquisition at the sub-word, word, or multi-word level. Few models assign structure or meaning to an entire utterance (or discourse) although many papers suggest that a presented model could be (easily) scaled-up – a worthwhile direction for future research. It is also worth remarking on the fact that articles addressing formal learning issues (e.g., PAC learning, identification in the limit, grammar induction, etc.) or that incorporate formalisms from mainstream computational linguistics (e.g., any of the many variants of probabilistic grammars) are underrepresented (the workshop contains one such). Future meetings along the lines of this workshop might benefit from attracting research efforts related to these approaches.

I would sincerely like to thank the program committee for above-and-beyond effort given the tight timetable, the diversity of the papers, and the several frustrating problems caused by spam-blockers; the workshop assistants who were a tremendous help with collating the reviews, organizing the articles for the proceedings, dealing with email and designing the conference web site; and, finally, the members of the COLING-2004 Workshop Program Committee, who were extremely helpful (and patient) on more than one occasion.

William Gregory Sakas
New York City
June 2004

INVITED SPEAKERS:

Walter Daelemans	(University of Antwerp, Belgium and Tilburg University, the Netherlands)
B. Elan Dresher	(University of Toronto, Canada)
Jerome A. Feldman	(University of California at Berkeley, USA)[1]
Charles Yang	(Yale University, USA)

ORGANIZER:

William Gregory Sakas	(City University of New York, USA)

PROGRAM COMMITTEE:

Robert Berwick	(MIT, USA)
Antal van den Bosch	(Tilburg University, the Netherlands)
Ted Briscoe	(University of Cambridge, UK)
Damir Cavar	(Indiana University, USA)
Morten H. Christiansen	(Cornell University, USA)
Stephen Clark	(University of Edinburgh, UK)
James Cussens	(University of York, UK)
Walter Daelemans	(University of Antwerp, Belgium and Tilburg University, the Netherlands)
Jeffrey Elman	(University of California, San Diego, USA)
Gerard Kempen	(Leiden University, the Netherlands and the Max Planck Institute, Nijmegen)
Vincenzo Lombardo	(University of Torino, Italy)
Larry Moss	(University of Indiana, USA)
Miles Osborne	(University of Edinburgh, UK)
Dan Roth	(University of Illinois at Urbana-Champaign, USA)
Ivan Sag	(Stanford University, USA)
Jeffrey Siskind	(Purdue University, USA)
Mark Steedman	(University of Edinburgh, UK)
Menno van Zaanen	(Tilburg University, the Netherlands)
Charles Yang	(Yale University, USA)

[1] Not listed in printed version of proceedings.

WORKSHOP ASSISTANTS:

Xuan Nga Cao (City University of New York, USA)
Mari Fujimoto (City University of New York, USA)
Lydiya Tornyova (City University of New York, USA)

SPONSOR:

The 20th International Conference on Computational Linguistics

FURTHER INFORMATION:

William Gregory Sakas
Ph.D. Programs in Linguistics and Computer Science
Department of Computer Science, North Bldg1008
Hunter College, City University of New York
695 Park Ave
New York, NY 10021
USA

email: sakas@hunter.cuny.edu
psycho.comp@hunter.cuny.edu

WWW: http://www.colag.cs.hunter.cuny.edu/psychocomp

Schedule of Presentations

8:40 *Introduction,* William G. Sakas

Morphology, Speech Segmentation

9:10 *On Statistical Parameter Setting,* Damir Cavar, Joshua Herring, Toshikazu Ikuta, Paul Rodrigues and Giancarlo Schrementi

9:35 *Combining Utterance-Boundary and Predictability Approaches to Speech Segmentation,* Aris Xanthos

10:00 Short Break

Phonology

10:10 Invited Talk: *On the Acquisition of Phonological Representations*, B. Elan Dresher

10:50 Coffee Break

Child Production

11:20 *A Computational Model of Emergent Simple Syntax: Supporting the Natural Transition from the One-Word Stage to the Two-Word Stage,* Kris Jack, Chris Reed and Annalu Waller

11:45 *Modelling Syntactic Development in a Cross-Linguistic Context,* Fernand Gobet, Daniel Freudenthal and Julian M. Pine

Connectionist

12:10 *On a Possible Role for Pronouns in the Acquisition of Verbs,* Aarre Laakso and Linda Smith

12:35 *Modelling Atypical Syntax Processing,* Michael S. C. Thomas and Martin Redington

13:00 Lunch

Constructionist

14:00 *Putting Meaning into Grammar Learning,* Nancy Chang

14:25 *Some Tests of an Unsupervised Model of Language Acquisition,* Bo Pedersen, Shimon Edelman, Zach Solan, David Horn and Eytan Ruppin

Principles and Parameters

14:50 *A Quantitative Evaluation of Naturalistic Models of Language Acquisition; the Efficiency of the Triggering Learning Algorithm Compared to a Categorial Grammar Learner,* Paula Buttery

15:15 *Grammatical Inference and First Language Acquisition,* Alexander Clark

15:40 Coffee Break

Learning Biases

16:10 Invited Panel: *Learning Biases in Language Acquisition,* Walter Daelemans, Jerome Feldman and Charles D. Yang

17:40 Closing remarks and discussion

Reserve paper: *A Developmental Model of Syntax Acquisition in the Construction Grammar Framework with Cross-Linguistic Validation in English and Japanese,* Peter Ford Dominey and Toshio Inui

Table of Contents