Project no.: **027657**

Project full title: **Perception, Action & Cognition through learning of Object-Action Complexes**

Project Acronym: **PACO-PLUS**

Deliverable no.: D5.2.2

Title of the deliverable: Grammar Induction

Contractual Date of Delivery to the CEC:	31 July 2008	
Actual Date of Delivery to the CEC:	12 Sept 2008	
Organisation name of lead contractor for this deliverable:	UEDIN	
Author(s): Mark Steedman, Ronald Petrick, and Christopher Geib		
Participant(s): UEDIN		
Work package contributing to the deliverable:	WP5	
Nature:	R	
Version:	Final	
Total number of pages:	42	
Start date of project:	1st Feb. 2006	**Duration:** 48 month

Project co-funded by the European Commission within the Sixth Framework Programme (2002–2006)

Dissemination Level

PU	Public	**X**
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Abstract:

The core focus of workpackage WP5.2 is to link the non-linguistic Object-Action Complex (OAC)-based conceptual representation developed under the PACO-PLUS project to language via a universal Language Acquisition Algorithm, and for the system to deploy the learned grammar in understanding and generating purposeful dialog with a human user.
As with human children, the conceptual representation that our systems induce from interaction with the world via low-level continuous control systems, such as the SDU robot/vision system in WP4.1, are language-independent. The langage acquisition algorithm must therefore be capable of learning the syntax of any human language from exposure to utterances pairing such conceptual representations with the appropriate sentence in that language, with the conceptual representation providing the semantics. Different languages partition that conceptual content into syntactic units such as word-meaning pairs in different ways. So our learning algorithm must consider all such partitions.
Since the problem of recognizing and planning dialog acts is an instance of the more general problem of planning with incomplete information and information-gathering actions, the OAC-based representation and the PKS planner developed under WP5 to support ordinary action planning, and reported on in D5.1.2, are immediately applicable to dialog planning, with objects generalizing to include agents, actions to include speech acts, and states to agents' knowledge states. The associated deliverable D5.1.2 describes the planner in greater detail.

Keyword list: Grammar Induction, Planning Dialog, Grounded Language Acquisition,

Table of Contents

1. Executive Summary 5
1.1 The Place of Language in the paco-plus Project 5
1.2 Language Acquisition Grounded in Automatically Induced Plans and Action Representations 5
1.3 Planning Dialog Acts 7
2. Publications Associated with D5.2.2 9
References 9
A. The Statistical Problem of Language acquisition 11
B. Planning Dialog Actions 35

1. Executive Summary

The core focus of workpackage WP5.2 is to link the non-linguistic Object-Action Complex (OAC)-based conceptual representation developed under the PACO-PLUS project to language via a universal Language Acquisition Algorithm, and for the system to deploy the learned grammar in understanding and generating purposeful dialog with a human user.

1.1 The Place of Language in the PACO-PLUS Project

As the proposal and Annex make clear, the role of language in the PACO-PLUS project is not primarily to act as a real-time user interface to the various robot platforms involved. Since the repertory of high-level actions, plans, and goals of the platforms will remain quite restricted, commercial speech recognition treating the identification of the user's utterances as a finite classification problem, encoding those states and actions. is always going to be adequate, and much faster and more reliable than full blown syntactic analysis and semantic interpretation, especially in the face of the high word error rates that can be expected from state-of-the-art speech recognition used as an input for parsing.

The place of language in the PACO-PLUS project is, rather, a theoretical investigation into the nature of language itself, and its ontogeny in human child-language acquirers in prelinguistic sensory-motor cognition, planning, and the Object-Action Complex (OAC) based knowledge representation developed elsewhere in the project. While we apply this theory to an artificially constructed corpus of utterances in the robot domain, a substantial emphasis on human language acquisition is involved, and the future deliverable D5.2.3 as specified in the Detailed Implementation Plan for M25-42 is principally concerned with real data of child-directed speech, and its relation to the differently-grounded artificial corpus. A substantial amount of work has already been done in the period up to M30 on transforming the dependency-annotated part of the CHILDES corpus into a quasi-semantic representation for this purpose, and will be reported on next year.

1.2 Language Acquisition Grounded in Automatically Induced Plans and Action Representations

As with human children, the conceptual representation that our systems induce from interaction with the world via low-level continuous control systems, such as the SDU robot/vision system in WP4.1, are language-independent. The language acquisition algorithm must, therefore, like a human child, be capable of learning the syntax of any human language from exposure to utterances pairing such conceptual representations with the appropriate sentence in that language, with the conceptual representation determining the semantics. Different languages partition that conceptual content into syntactic units such as word-meaning pairs in different ways. So our learning algorithm must consider all such partitions.

For example, French and English differ systematically in the way they lexicalize causative actions as verbs. English makes profligate use of various particles, adverbs, and transitive constructions to make "undirected activity" verbs like "swim" into "directed activity" verbs like "swim across the Rio Grande" and "swim the English Channel", as in (1a). However, French lacks such causatives, and has to use a periphrastic manner adverbial with a lexicalized directed activity verb such as *aller* or *traverser*, as in (b):

(1)
a. The children swam across the lake.
b. Les enfants ont traversé le lac à la nage.

It is reasonable to assume that the child learning either language has access to a meaning representation that is closer to the more articulated French version than the English—say for the sake of illustration, something

like the following:

(2) $past'(go'(across'(the'lake'))(the'children')(by'swimming'))$

Both French and English learners must respectively entertain the hypothesis not only that *swam* or *ont traversé* means "went across" (which is correct for the latter, but incorrect for the former), but also that they mean "went by swimming" (which is correct for the former, but not for the latter).

Our robots raise the same problem, even in their very restricted knowledge domains. For example, the Odense hand-eye system and the PKS planner are capable of forming PKS plans to achieve a state in which all the cups currently on the table are in a box, using symbolically-represented operators that can, at least in principle, be learned by associating consistent changes in the world contingent upon actions like grasping and moving applied to objects like cups (D5.1.2, Mourão *et al.* 2008b,a), and which together with plans, support a simple linguistic semantics.. Consider the problem of learning corresponding fragments of French and English sufficient to converse about this limited domain. To speed up the discussion, let us assume that the lexical item for cup has already been learned, (although even this step is non-trivial, cross-linguistically speaking).

Suppose we first want the system to learn the English verb "grasp". Unlike a child, for the Odense system the grasping action is currently a monolithic concept, since the latter only has one effector, so that the programmers (or the machine learning system described in D5.1.2) have no reason to distinguish prehension with the gripper from any other kind.

At first glance, the English system has no interesting language learning problem: it seems as if it just has to associate the word "grasp" with this plan-level sensory motor primitive. The language acquisition algorithm described in appendix A will allow it to do this on the basis of commentaries like the following:

(3) You GRASP the cup!

However, French users are more likely to use the correct equivalent of English "grasp", namely "prendre à la main" as the correct but artificial "prendre," as in the following:

(4) Tu prends la tasse à la MAIN!

Our algorithm does not allows the language learner to hypothesise discontinuous constituents like "prends X à la main." It will instead learn that "prendre" means "grasp", and treat "à la main" as a semantically vacuous modifier. This tells us, rather obviously, that the human child is differently grounded to the Odense robot. Either we have to live with that, or we have to "cheat" by building into the grasp concept the fact that it is structured into prehension and effector components. In the latter case we must say goodbye to any possibility of applying machine learning to pure unsupervised interaction with the world, of the kind explored in D5.1.2. We choose the former course: in the end, even in conversing with other human beings, we have to deal with the fact of private language, differently grounded in differentially enabled individuals, as Landau and Gleitman 1985 showed for blind children.

Suppose that we next want the system to learn that the verb-particle compound "put away" corresponds to the plan of making some thing or set of things not on the table. In the current implementation of the Odense system, there is only one place apart from the table, namely the box, so the system is not in a position to semantically distinguish any of the following utterances:

(5)
a. You put your cups in the box!
b. You put away your cups!
c. You clear the table of cups!

In particular, it will learn that one meaning for "put away X" is the plan "while there is an X on the table, put it in the box". Again this is an overly specific private language in comparison to human language, but as with humans, the parsing model will disambiguate the "put way" plan that is appropriate to cups from that for other objects, once it has some other objects and places to think about. The problem of generalizing from different instances of putting away to a general plan (and corresponding semantic primitive) that can be applied to novel objects remains to be dealt with at a later stage.

In the case of French, the corresponding commentary is the following, in which "ranger" is directly equivalent to "put away".

(6) Tu ranges tes tasses!

The operation of our language acquisition is to be contrasted with those in other research projects concerned with the induction of affordance-like action concepts and/or the relation of action to language. Most such work either seeks to induce sensory motor action representations without primary regard to the relation to the symbolic level related to planning and language, such as that under the EU MACS and Robocub projects (Sahin *et al.* 2007, citealtOrab:05), or assumes a conceptual representation that is somewhat arbitrarily modeled on a presumed semantics for a particular language, for which it either imposes or learns a pairing between a grammar and the conceptual relation, as in the EU JAST and COSY projects (Foster *et al.* 2006, Jacobsson *et al.* 2008, and the work of Deb Roy's group, which is in other respects close in spirit to the present work (Roy 2005; Gorniak and Roy 2005, 2007). While our results will inevitable be smaller-scale than those systems, and for the restricted applications available in PACO-PLUS , it will remain no more than a theoretically interesting but practically cumbersome alternative to brute force speech recognition treating the limited variety of utterances that the robots encompass as whole words, we believe that our approach raises deep and fundamental issues about the nature of grounded language acquisition that are of lasting interest.

1.3 Planning Dialog Acts

Since the problem of recognizing and planning dialog acts is an instance of the more general problem of planning with incomplete information and information-gathering actions, the OAC-based representation and the PKS planner developed under WP5 to support ordinary action planning, and reported on in D5.1.2, are immediately applicable to dialog planning, with objects generalizing to include agents, actions to include speech acts, and states to agents' knowledge states. The associated deliverable D5.1.2 describes the planner in greater detail.

We have attached a number of additional documents to this deliverable that highlight the full generality of the language acquisition algorithm summarized above, and the potential for PKS planning as a basis for flexible dialog applications. Both go very considerably beyond the, in linguistic terms, very narrow perspective of the PACO-PLUS project. We make no apology for this breadth. It is necessary in order to prove that our claims for the possibility of full human language acquisition and full human language use on the basis of conceptual representations grounded in action in the social world are more than a fashionable metaphor, and scale to the full complexity of human langauge and its use.

The documents describe a statistical-model-based language acquisition algorithm and the specifically communicative knowledge-base that would in principle allow any robot platform that supports PKS-style planning (of which the PACO-PLUS project currently includes two) the capacity to acquire and use any human language to achieve goals in its domain. More generally, these components provide the infrastructure needed to support much longer term objectives of language and communication. Here we briefly sketch the relation of each paper to this workpackage and deliverable, and make links to the specific contributions of each paper.

[A] *(to be submitted)* This paper defines the langauge acquisition algorithm that induces a grammar from utterance-meaning pairs via an incrementally-built parsing model. A number of further ramifications of the language acquisition process are discussed, including the pervasive influence in the later stages of "syntactic bootstrapping.

[B] *(Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue, Antwerp)* This paper analyzes the problem of planning dialog acts in terms of the PKS planner and the problem of planning with sensing actions. Sensing actions and speech acts complicate planning by threatening to engender potentially infinite state spaces. The PKS planner uses program variables (Etzioni *et al.* 1992 to overcome both problems.

Together, these papers report a number of significant developments:

- A number of theoretical devices that have been proposed in the linguistic literature, and which are computationally problematic, are simply redundant, according to our proposal. These include the system itself of "parameters" proposed by Chomsky, the notion of "triggers" for setting parameters, and the "subset principle" that makes the notion of triggers necessary.
- An axiomatization of a speech act knowledge domain is proposed which exploits PKS's representation of knowledge-gathering and sensing actions to define reduced search spaces for dialog plans which may allow practical planning with richer representations of dialog states that are allowed by the state-of-the-art POMDP representation for dialog managers.
- An analysis of the phenomenon of "indirect" speech acts, including the "conversational implicatures" described by Grice, which bypasses his apparatus of recognition of intention and invocations of Maxims and Principles of Cooperation, simplifyiong the problem of dialog
- The work described in this report and its companion D5.1.2 in interaction with the other workpackages of PACO-PLUS , provides a complete theoretical path from continuous low-level representations to human-scale language acquisition and dialog-level representations.

A number of questions remain open at the time of this report and constitute further work.

- We have yet to apply the language-learning to the problem of learning from a corpus of human-to-machine utterances of the type that could be supported by the Odense and Karslruhe platforms. This is solely due to the fact that the variety of actions that can be represented for those platforms at the lecel of the planner (and hence at the level needed to support truly grounded language acquisition) continues to be very small. Further object and action classes are under development in collaboration with Odense under WP4. UEDIN has also begun integrating its control, communication, and planning mechanisms on the ARMAR robot platform as part of WP1.
- We have yet to identify an application domain in which the theoretical advantages of our dialog planner can be shown to be of practical benefit in competition or in interactions with POMDP techniques. One problem is the lack of well defined benchmarks. We are currently evaqluating a modem-user troubleshooting domain (Boye 2007)
- Since nondeterminacy will undoubtedly arise as the result of perception and action at the discourse level, we are studying how best to utilise such information in speech-act planning. The work will be informed by work disussed in D5.1.2
- Although the theoretical work required to extend PKS to support dialogue planning of the form described here is complete, the implementation of these extensions (Milestone 5.1.1) is only partially complete at the time of reporting.

Besides the connections to WP1, WP4, and WP5 already mentioned, this workpackage also has interactions with other workpackages including WP2, WP3, and WP7.

2. Publications Associated with D5.2.2

[A] **The Statistical Problem of Language Acquisition**
Mark Steedman, Tom Kwiatkowski, and Julia Hockenmaier.
To be submitted. Part of this draft presented at ICL 2008 Seoul Korea as "The Computational Problem of Language Acquisition"

> **Abstract:** From the point of view of strongly lexicalized theories of grammar, the task that faces the child in the earliest stages of language acquisition is simply that of learning a language-specific lexicon on the basis of exposure to (probably contextually ambiguous, possibly somewhat noisy) sentence-meaning pairs, given a universal grammatical "projection principle", and a similarly universal functional mapping from lexical syntactic types to semantic types in a universal language of logical form.
>
> The paper argues that, under these assumptions, a very simple statistical model allows children to arrive at a target lexicon without navigation of subset principles, or attention to any attendant notion of trigger other than that of a "reasonably short sentence in a reasonably understandable situation drawn from a reasonably representative sample". The model explains the general pattern of errors that are found in elicitation experiments. The linguistic notion of "parameter" appears to be entirely redundant to this process.

[B] **Planning Dialog Actions**
Mark Steedman and Ron Petrick
Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue, Antwerp, Sept. 2007, 265-272.

> **Abstract:** The problem of planning dialog moves can be viewed as an instance of the more general AI problem of planning with incomplete information and sensing. Sensing actions complicate the planning process since such actions engender potentially infinite state spaces. We adapt the Linear Dynamic Event Calculus (LDEC) to the representation of dialog acts using insights from the PKS planner, and show how this formalism can be applied to the problem of planning mixed-initiative collaborative discourse.

References

Boye, Johan, 2007. "Dialogue Management for Automatic Troubleshooting and other Problem-Solving Applications." In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*. Antwerp: ACL, 247–255.

Etzioni, Oren, Hanks, Steve, Weld, Daniel, Draper, Denise, Lesh, Neal, and Williamson, Mike, 1992. "An Approach to Planning with Incomplete Information." In *Proceedings of the 2nd International Conference on Knowledge Representation and Reasoning (KRR-2)*. 115–125.

Foster, Mary Ellen, By, Tomas, Rickert, Markus, and Knoll, Alois, 2006. "Human-Robot Dialogue for Joint Construction Tasks." In *Proceedings of the 8th International Conference on Multimodal Interfaces*. Banff, Alberta: ACM, 68 – 71.

Gorniak, Peter and Roy, Deb, 2005. "Probabilistic Grounding of Situated Speech Using Plan Recognition and Reference Resolution." In *Proceedings of the 7th International Conference on Multimodal Interfaces (ICMI 2005)*. ACM.

Gorniak, Peter and Roy, Deb, 2007. "Situated Language Understanding as Filtering Perceived Affordances." *Cognitive Science* 31. to appear.

Jacobsson, Henrik, Hawes, Nick, Kruijff, Geert-Jan, and Wyatt, Jeremy, 2008. "Cross-Modal Binding in Information Processing Architectures." In *Proceedings of the 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI-08)*. New York: ACM, 81–88.

Landau, Barbara and Gleitman, Lila, 1985. *Language and Experience: Evidence from the Blind Child*. Cambridge MA: Harvard University Press.

Mourão, Kira, Petrick, Ron, and Steedman, Mark, 2008a. "On Learning the Dynamics of Planning Domains." In *Proceedings of 3rd Workshop on Ontology Learning and Population (OLP3), ECAI 2008*. University of Patras. to appear.

Mourão, Kira, Petrick, Ron, and Steedman, Mark, 2008b. "Using Kernel Perceptrons to Learn Action Effects for Planning." In *Proceedings of 3rd International Conference on Cognitive Systems (CogSys 2008)*. University of Karlsruhe. six pages.

Roy, Deb, 2005. "Semiotic Schemas: A Framework for Grounding Language in Action and Perception." *Artificial Intelligence* 167:170–205.

Sahin, E., Cakmak, M., Dogar, M.R., Ugur, E., and Ucoluk, G., 2007. "To Afford or not to Afford: A New Formalization of Affordances towards Affordance-Based Robot Control." *Adaptive Behavior* 15:447–472.

Appendix A

The Statistical Problem of Language Acquisition

Mark Steedman, Tom Kwiatkowski, and Julia Hockenmaier

September 12, 2008

Abstract

From the point of view of strongly lexicalized theories of grammar, the task that faces the child in the earliest stages of language acquisition is simply that of learning a language-specific lexicon on the basis of exposure to (probably contextually ambiguous, possibly somewhat noisy) sentence-meaning pairs, given a universal grammatical "projection principle", and a similarly universal functional mapping from lexical syntactic types to semantic types in a universal language of logical form.

The paper argues that, under these assumptions, a very simple statistical model allows children to arrive at a target lexicon without navigation of subset principles, or attention to any attendant notion of trigger other than that of a "reasonably short sentence in a reasonably understandable situation drawn from a reasonably representative sample". The model explains the general pattern of errors that are found in elicitation experiments. The linguistic notion of "parameter" appears to be entirely redundant to this process.

1 INTRODUCTION

It is widely agreed that, in learning such basic aspects of language-specific grammar as which words of the language are the verbs and which the nouns, and in what linear spatio-temporal order(s) the two may occur, children must have access to something more than mere strings of words constituting a subset of the legal sentences of the languages.

This agreement is based in part on observation of the extreme rapidity with which language acquisition proceeds, and the absence of negative data in the input to the child. While it is theoretically possible, using probabilistic models and unsupervised machine learning, to approximate grammars of linguistically relevant classes to any desired degree of accuracy (Horning (1969)), the computational costs of such learning for realistic grammars are prohibitive, and there has been little success so far in practical unsupervised induction of natural language grammars from positive data alone.

There is much less agreement concerning the actual nature of the "something more" that the child brings to the task. It is sometimes referred to as "Universal Grammar", and as such is sometimes talked about in exclusively syntactic terms, as in the "parameter-setting" account of language acquisition of Hyams (1986) and much subsequent work. According to this account, a homunculus "flips switches" corresponding to syntactic parameters such as head-finality and *pro*-drop until the "universal grammar engine" uniquely specifies the language *modulo* its lexicon, in a process that has been likened to a game of Twenty-Questions (Yang 2006:Ch.7).[1]

However, such accounts raise as many questions as they answer about the mechanism by which such learning could proceed. In particular, the specific inventory of parameters that this universal machine embodies, the way in which the very large search spaces engendered by even quite small sets of binary independent parameter can be effectively explored (Clark & Roberts 1993;Fodor & Sakas 2005), and the aspects of the data that "trigger" their setting (Gibson & Wexler 1995; Fodor 1998) remain rather unclear. This is sometimes referred to as the "logical problem of language acquisition."

There is something deeply appealing in the idea that the process of language learning proceeds by entertaining all possible grammars, and eliminating all alternatives but one, because that is pretty much what the child's developmental behavior looks like. In particular, Crain & Thornton (1998) and their students have shown (using ingeniously forced elicitations) that learning is characterized by great initial variation in productions for any given construction, apparently covering alternatives characteristic of many other languages, followed by abrupt transitions to stable adherence to the correct form for the target language.

[1] One is uneasily reminded of the warnings of Newell (1973) in a different context, concerning the likely outcome of "playing Twenty-Questions with nature."

Yang (2002) offers a probabilistic account of this process in terms of classical Mathematical Learning Theory. While Thornton & Tesan (2006) argue that changes they observe are too abrupt and switch-like to support that particular model, probabilistic models in general are capable of approximating catastrophic, switch-like behavior, so they should not be ruled out.

The present paper, following work by Siskind (1996), Villavicencio (2002), and Zettlemoyer et al. (2005), shows that a very simple statistical model and learning algorithm makes the notion of parameter-setting entirely redundant. The only notion of trigger that it requires is the notion "reasonably short sentence with an independently accessible meaning". The only notion of language specific grammar it needs is the lexicon for the language. The only notion of universal grammar that it needs is a universal mapping from each semantic type to the possible lexical syntactic types, together with a universal machine for merging or projecting lexical types and their meaning representations onto grammatical derivations. The former element is the origin of Chomsky's (1965) "substantive" universals concerning linguistic categories, while the latter is the origin of "formal" universals concerning syntactic projection.

2 Semantically Grounded Grammar Acquisition

The most plausible source for substantive universals is a universal *semantics*, broadly construed, in the form of structured meaning representations, closely related to the conceptual representations that enable the child's cognitive understanding of the world, to which the child already has access as language acquisition begins (Chomsky 1965:27-30; Chomsky 1995:54-55), and to which syntactic forms are rather directly attached, drastically limiting the search space.

To say this much is not very helpful in psychological or linguistic terms, since (as Chomsky never tires of pointing out) linguists don't know very much about how to articulate the semantics. One of the problems that they face is that human semantics is greatly, perhaps even mainly, concerned with highly dynamic interpersonal, social, and intentional content, of a kind that is deeply embedded in physical and interpersonal interaction with the world, and distinctly under-represented in formal linguistic theories of semantics Tomasello (1999).

However, the child doesn't *need* to articulate such a semantics. They just need to label it with linguistic categories, so our theories need to represent it somehow. As a temporary stopgap we'll use terms of the lambda calculus, and defer the question of what a more psychologically realistic human semantics might actually look like till section 8.

This approach makes the child's problem resemble that of treebank grammar induction for wide coverage parsing (Collins 1997; Charniak 2000; Hockenmaier & Steedman 2002), where sentences hand-annotated with syntactic trees are used to derive a grammar and a statistical parsing model. However, the child's task is a little harder. First, they have to induce the grammar from strings paired with *unordered logical forms*, rather than language-specific ordered derivation trees. That is, they have to work out *which word(s) go with which element(s) of logical form*, as well as the directionality of the syntactic categories (which are otherwise universally determined by the semantic types of the latter). Second, while they do not seem to have to deal with a greater amount of error than is found in the Penn WSJ treebank (McWhinnie 2005), they may need to deal with *situations which support a number of logical forms*. Third, they need to be able to recover from temporary *wrong lexical assignments*. Fourth, they need to tolerate *lexical ambiguity*.

3 Previous Work

Siskind (1995, 1996),Thompson & Mooney (2003), Villavicencio (2002), and Zettlemoyer et al. (2005) offer computational models of the process of inducing a grammar from string-meaning pairs, the latter two explicitly using CCG.[2]

Siskind and Villavicencio make strong assumptions about the association of words with elements of logical form. Both make similarly strong assumptions about universally available parametrically specified rule- or category- types, the latter assuming a type hierarchy. Both deal with noise and homonymy probabilistically.

[2]This approach is to be contrasted with the method of Kanazawa (1998) for learning *k*-valued categorial grammars, and the related work of Buttery (2006), in which all possible structures are assigned to strings and labeled according to the algorithm of Buszkowski & Penn 1990, and lexical hypotheses are distinguished on non-semantic criteria. See also the related approach of Osborne & Briscoe (1997), which uses a minimum description length (MDL) criterion.

Both do the learning in two stages, first associating logical forms with words, then inducing phrase structure rules (Siskind) or directional CCG categories (Villavicencio).

However, there is no necessity to separate the two processes of associating meaning and syntactic type. Zettlemoyer and Collins (2005) combine the two in a single pass CCG induction algorithm. Crucially, their algorithm allows *any contiguous substring* of the sentence to be a lexical item, so that for the given logical form, the learner has to search the cross-product of the substring powerset of the string with the set of pairs of legal categories of the substructure powerset of the logical form, as in the example (9) below, for categories that yield combinatory derivations that yield the correct logical form. Learning is via a log-linear model using lexical entries as features and gradient descent on their weights, iterating over successive sentences of a corpus of sentence-logical form pairs.

The algorithm as presented in 2005 learns only a very small rather unambiguous fragment of English, hand-labeled with uniquely identified database queries as logical forms, and an English specific inventory of possible syntactic category types in lieu of Universal Grammar, without the involvement of a parsing model. However, Siskind's and Villavicencio's results already tell us that the algorithm should work with multiple candidate logical forms. Similarly, their results show that a universal set of category types can be used without overwhelming the learner.

All of these models depend on availability to the learner of short sentences paired with logical forms, since complexity is determined by a cross-product of powersets both of which are exponential in sentence length. The use of statistical models is also crucial in handling this complexity.

Because it allows multiword elements (MWE) to be lexical entries, Zettlemeyer and Collins' program avoids the problem that two words which consistently collocate, like *want* and *to* fail to reveal which of them means *want′* and which means *to′*. They can be learned as a single item *want to*. So can idioms and multi-word expressions like "buy the farm," and "take advantage of"

As with Siskind's version, lexical items can have complex meanings—corresponding for example to causatives, whose availability may differ (*swim across* vs. *traverser à la nâge*, *put away* vs. *ranger*, etc.) across languages. No notion of trigger distinct from that of "reasonably simple string-meaning pair" is necessary.

It is possible to use the statistics of the lexicon itself to implicitly represent "parameters" such as verb-finality, via incrementally adjusted prior probabilities on the members of the set of universally available category types.

4 The Proposal

We will assume as a theory of grammar a version of Combinatory Categorial Grammar (CCG, Steedman 2000b; Steedman & Baldridge 2006) in which all language-specific information resides in the lexicon, and a universal set of combinatory rules including functional composition and a case-like lexicalized operation of type-raising, as well as function application, projects strings of lexical items onto sentence-meaning pairs.

The primary task that the child faces is to learn the categorial lexicon on the basis of exposure to (possibly situationally ambiguous, possibly somewhat noisy) sentence-meaning pairs, given this universal combinatory projection principle, and a mapping from semantic types to the set of all universally available lexical syntactic types. To do this soundly and efficiently, the primary desideratum for such a system is that information gained in earlier stages of learning should propagate to unseen items encountered at a later stage. Thus if every word or rule except one in a novel sentence has been seen with the category required for its analysis, leaving only one possibility for the unseen word, that information should be reflected in a high probability for that value.

4.1 The Uses of Statistical Models

Although it is assumed here that the child learns (at least initially) from sentences whose meanings it knows, this is clearly not the situation that speakers generally face – typically, the meaning of a sentence is not previously known to the listener, otherwise there would be no need for communication. But because natural language is ambiguous, the child also faces the task of acquiring a model which allows it to disambiguate (i.e to identify the most likely intended meaning) of new sentences. And because the child will also want to produce language, it requires a model that allows it to choose strings of words that express the meaning it wants to convey. However, under the assumption that grammar models are probabilistic (and that the child is capable of elementary operations of probability theory, such as Bayes' Rule), there is no need to

stipulate two (or possibly three) distinct grammar models for each of these tasks. We will, in fact assume that a single model underlies parsing/comprehension, generation/production and language acquisition. This model $\mathbf{M} = \mathbf{P}(\mathbf{I}, \mathbf{S}, \mathbf{D})$ defines a joint probability distribution over semantic interpretations I, surface strings or sentences S, and (CCG) derivations D. Such distributions are commonly used in statistical parsing (where the interpretation I is either ignored or approximated with a word-word dependency structure). Because the set of derivations (and sentences) is infinite, the distribution $P(I,S,D)$ cannot be estimated directly, but is standardly assumed to be generated by an underlying stochastic process which, for context-free grammars and CCGs, mimicks top-down derivations that are all rooted in the unique start symbol of the grammar.

We will assume that $P(D,I,S)$ is a generative model for an (exhaustive) parser, rather than the discriminative model of Zettlemoyer et al.. One advantage of generative models besides their closeness to competence grammar is that we can invert the parsing model to define the probability of an utterance given a meaning. Since the way the model is derived is not immediately relevant, and the initial model defines everything as equiprobable (or otherwise determined by UG), we defer further discussion.

Basic idea of a generative model

During language acquisition, the model $P(D,I,S)$ can be used to calculate a distribution of possible derivations D and interpretations I for the input sentence S:

$$P(I,D|S) = \frac{P(I,D,S)}{\sum_{I,D} P(I,D,S)} \tag{1}$$

In a fully supervised learning algorithm (i.e. if the learner is exposed to a data set $X = \{\langle I,S,D\rangle\}$, consisting of sentence-derivation-interpretation tuples), any distribution $P(A|B)$ can be computed by simply counting the number of occurrences, or the frequency f, of A and B:

$$P(A|B) = \frac{f(A,B)}{\sum_{A'} f(A',B)} \tag{2}$$

Such relative frequencies of A given B are well known to yield a maximum-likelihood-estimate of $P(A|B)$, i.e. a distribution which assigns the highest probability to the data set X.

However, language acquisition is not fully supervised. The child is faced with a great deal of uncertainty regarding the way the sentence is split into words, the meaning of the sentence, and its possible derivations. To simplify the presentation, we will initially assume that the child learns from sentences whose interpretation it knows unambiguously.

Our model $P(I,D,S)$ (or rather, its component distributions) defines a set of expected frequencies of events such as word-category pairs, or particular rule instantiations. We assume that the child (and, indeed, the adult) updates its model with these relative expected frequencies after it has been exposed to a new sentence S. Such a learning procedure corresponds to an incremental, or online, version of the Expectation-Maximization (EM) algorithm. It is well known that EM is not in general guaranteed to find the optimal solution, and can get stuck in local minima. However, because we assume that (at least until the grammar is relatively stable) only sentences with a known interpretation are used in this process, our proposal corresponds to (an online version of) the semi-supervised EM algorithm of Pereira & Schabes 1992.

For example, in a corpus of sentences S_i, each with a number of interpretations I_j, each of which has an number of derivations D_k, the expected frequency f_{exp} of a lexical entry $\langle \phi, \sigma, \lambda\rangle$ for a word with phonology ϕ, syntactic category σ and logical form λ (e.g.$\langle$`doggies`$,N,dog'\rangle$) occuring $n_{i,j,k}$ times per derivation is given by:

$$fexp_{lex}(\langle \phi, \sigma, \lambda\rangle) = \sum_i \sum_j P(I_j|S_i) \sum_k P(D_k|I_j,S_i) \cdot n_{i,j,k}(\langle \phi, \sigma, \lambda\rangle) \tag{3}$$

The probability of a lexical entry can be defined in terms of its expected frequency (3) as:

$$P_{lex}(\langle \phi, \sigma, \lambda\rangle) = \frac{f_{exp}(\langle \phi, \sigma, \lambda\rangle)}{\sum_i f_{exp}(\langle \phi, \sigma, \lambda\rangle_i)} \tag{4}$$

(there might be more than one syntactic type σ per logical form λ_τ)

Similarly, the conditional probability of uttering a word ϕ, such as "more" or "doggies", given a logical form λ_τ, such as $more'_{((e,t),e)}$, can be obtained as follows:

$$P(\phi|\lambda_\tau) = \frac{\sum_i f_{exp}(\phi, \sigma_i, \lambda_\tau)}{\sum_{j,k} f_{exp}(\phi_j, \sigma_k, \lambda_\tau)} \tag{5}$$

It will also be useful to know that the prior conditional probability of a syntactic type σ given a logical form λ_τ of semantic type τ can be obtained as follows:

(6) $$P(\sigma|\tau) = \frac{\sum_{i,k} f_{exp}(\phi_i, \sigma, \lambda_{\tau k})}{\sum_{i,j,k} f_{exp}(\phi_i, \sigma_j, \lambda_{\tau k})}$$

The course of language acquisition can then be described as follows.

4.2 The First Few Words

Consider an adult-accompanied child at Piagetian Stage VI who has yet to learn her first word of such a grammar. She encounters a dog, and shows great interest, but fails to show any evidence of having learned the word "doggie". Later, she encounters some *more dogs*, and exhibits wild excitement. The adult observes the child's reaction, and says "MORE DOGGIES!."

We can assume that the child has already learned some phonological regularities of the language, and in particular is in a position to consider the possibility that the utterance consists of more than one word (Mattys et al. 1999; Mattys & Jusczyk 2001).

We will further assume that the first thing the child does is to take the string-category-logical form triple $\langle Moredoggies := NP : more'doggies' \rangle$, and apply the rules of the grammar to it in the generative direction, to retrieve the final step of every derivation of that category possible under universal combinatory grammar. (This involves knowing the mapping between (known) semantic types and (unknown) legal CCG syntactic categories.) This procedure is then recursively applied to all of the pairs of possible daughters, terminating when either the whole substring is treated as a lexical entry or the logical form is an atom (in which case, the latter step is forced. This results in a "shared forest" of derivations that efficiently represents the set of all derivations licensed by UG, resembling the "chart" of standard parsing algorithms like CKY, and forming a basis for calculating frequencies of events for the statistical parsing model, according to the algorithm given below.[3]

The only two combinatory rules that have a non-function category as their result are the rules of function application:

(7) *Forward Application:*
$\langle \phi_l, X/Y, F \rangle \quad \langle \phi_r, Y, A \rangle \Rightarrow \langle \phi, X, FA \rangle$
backward Application:
$\langle \phi_l \rangle \quad \langle \phi_r, X/Y, F \rangle \Rightarrow \langle \phi, X, FA \rangle$

Since we know the value of all elements of the result, then if we know the universal syntactic types that correspond to the semantic types of F and A, we know all possible values of all elements of the left hand side. The utterance to hand, "More doggies", generates just three derivations, as follows:[4]

(8) a.
$$\frac{\dfrac{\text{MORE}}{NP/N : more'_{((e,t),e)}} \quad \dfrac{\text{DOGGIES}}{N : dogs'_{(e,t)}}}{NP : more'dogs'_e}\,{>} \quad !$$

b.
$$\frac{\dfrac{\text{MORE}}{N : dogs'_{(e,t)}} \quad \dfrac{\text{DOGGIES}}{NP\backslash N : more'_{((e,t),e)}}}{NP : more'dogs_e'}\,{<} \quad !$$

c.
$$\frac{\text{MORE DOGGIES}}{NP : more'dogs'_e} \quad !$$

[3]This is a departure from the related approach of Zettlemoyer et al. (2005), who assume that the child considers a larger set of lexical candidates onstructed by taking the cross-product of every non-empty substring ϕ of the utterance "More doggies!" with every connected typed subterm λ_τ of type τ the logical form $more'doggies'$, together with all syntactic categories σ_i that universal grammar allows for the semantic type τ of each such subterm. For the example to hand, this set would include certain potential categories, such as more:=$NP\backslash N : more'$ for which there is in fact no evidence from the sentence "more doggies!", and which the present algorithm will not generate. These spurious candidates are subsequently excluded by running a CCG parser over all possible sequences of lexical categories compatible with the string and detecting the fact that there is no derivation that they take part in.

[4]The example is simplified for exposition. The assumption that the child immediately considers the hypothesis that more is a determiner is particularly far-fetched, and will be reviewed later.

The following set of candidate lexical entries can be read off the three derivations in (8):

(9) The child's lexical candidates:

more:=	$\mathbf{NP/N} : \mathbf{more}'_{((e,t),e)}$
	$N : dogs'_{(e,t)}$
doggies:=	$NP\backslash N : more'_{((e,t),e)}$
	$\mathbf{N} : \mathbf{dogs}'_{(e,t)}$
more doggies:=	$NP : (more'dogs')_e$

The set (8) of derivations also defines a partial generative parsing model P. Since the rules in the set of derivations are highly interdependent, the naive PCFG which we begin by defining in terms of the set of productions involved in those derivations, and the conditional probabilities of those productions given the parent category involved in those derivations that rewrite preterminals, such as $N : dogs'$ as words such as "doggies" or (spuriously) "more", will not give a maximimum likelihood estimator. Rather than using a discriminative method, as in Zettlemoyer et al. (2005), we will eventually need to lexicalize the naive model P using *head word dependencies* (Collins 1997), in a way that we will come to later.

In the interests of psychological realism as well as computational economy, we propose an incremental algorithm that updates the model on the basis of each new sentence, rather than a batch model that recomputes the model for the entire corpus when each new sentence is encountered. The primary requirement for such a model is that *learned information about seen events in a derivation should influence the probabilities assigned to unseen events*. Thus, if the language only consists of sentences of the form "More X", and the hundredth sentence is "More erasers", where "erasers" is a previously unseen word, this sentence should not only make the learner a little more certain that "more" is a determiner meaning $more'$, and not a highly ambiguous word meaning $erasers'$, among other things like $dogs'$. It should also make them pretty sure that "erasers" is a noun, and *not* yet another determiner meaning $more'$.

We can define a model for a set of sentences S, each with a number of possible interpretations I, each giving rise to a set D of derivations in terms of a vector *fexp* of expected frequencies for each production p, defined as

$$(10)\quad fexp(p) = \sum_{s\in S}\sum_{i\in I} P(i|s)\sum_{d\in D} P(d|s,i).count(p,d) \text{ where} P(d|s,i) = \frac{P(d)}{\sum_{d\in D} P(d)}$$

We can view an estimator for an increment to the cumulative expected frequencies (10) based on observing the nth sentence as the sum of two weighted components, an "a posteriori" component, stemming from what we have already learned, and an "a priori" component defined by all possibilities allowed by universal grammar—thus:

$$(11)\quad \Delta fexp = \kappa \Delta fexp_{prior} + (1-\kappa)\Delta fexp_{posterior}$$

$Fexp_{posterior}$ for a given interpretation i for sentence s is defined as follows, where P is the model estimated so far.

$$(12)\quad fexp_{posterior}(p) = \sum_{i\in I} P(i|s)\sum_{d\in D} P(d|s,i).count(p,d)$$

$Fexp_{prior}$ is the expected frequency based on the present sentence and the possibilities of universal grammar alone. For simplicity we will assume the latter to be uniformly distributed, so that (10) reduces to the following, where $|D|$ is the number of derivations:

$$(13)\quad fexp_{prior}(p) = \sum_{d\in D} \frac{count(p,d)}{|D|}$$

Such a model, including the implicit changes in the κs over time, can be learned using the following incremental variant of the semi-supervised inside-outside (EM) algorithm (Dempster et al. 1977; Pereira & Schabes 1992; Neal & Hinton 1999).

Every new sentence s_n provides a set D_n of derivations parallel to (8), for which the following procedure is applied:

a. For each of a (possibly empty) set of *previously unseen productions* involved in some derivation in D_n, including those involving novel lexical entries, must be added to the model with cumulative *fexp* temporarily initialized to zero.

b. (E-step): For *all productions* involved in some derivation in D_n (including those in a), the cumulative *fexp* must be multiplied by $n-1$, incremented by $fexp_{prior}$, and divided by n.

c. (M-step): For *all productions* involved in some derivation in D_n (including those in a), a further increment of $\frac{fexp_{posterior} - fexp_{prior}}{n}$ (which may be negative) must be added to to the cumulative *fexp* . (I.e., the earlier increment based on $fexp_{prior}$ must be replaced by that based on $fexp_{posterior}$.)

Step b defines new values for the conditional probabilities for the rules in question, defining an intermediate model for calculating the a posteriori probabilities in step c. The further update c to the model defines the expected frequencies for the next cycle. The lexical probabilities for the relevant words in the lexicon given the new sentence can then be calculated using the model and definition (10), where $P(d|s,i)$ is the product of the probabilities of the productions it involves.

(14) $P(d|I,S) = \prod_{p \in d} P(p|parent) \prod_{LEX(p) \in d} P(\phi, \sigma|\lambda)$

This is just a probabilistic context-free grammar parser (PCFG). We actually need a head-dependency model (Collins 2003) of the kind described in the appendix.

The possibility of lexicalizing more than one element of the logical form in a single word means that the alternative derivations for a single logical form such as those in (8) for our running example and the first sentence "More doggies" may be of different lengths. Since generative models of the kind outlined above, based on the products of probabilities of rules, assign undue weight to short derivations, we must normalize the probabilities of lexical productions over the complexity of their logical forms. Thus, the probability $P(\phi.\sigma|\lambda)$ of the lexical productions in (14) is

(15) $P(\phi.\sigma|\lambda) = \prod_{m \subset \lambda} P(\phi, \sigma|m)$

For example, the probability of derivation (8c) is not a third, but is the conditional probability of "more dogs" given *more′dogs′* times that of "more dogs" given *more′*, times that of "more dogs" given *dogs′*—that is, $\frac{1}{3} \times \frac{1}{3} \times \frac{1}{3}$.

Thus, on the basis of the intermediate value $\frac{(0)fexp(0) + fexp_{prior}}{1}$, the relative conditional probabilities $p(d|i,s)$ of the three derivations (8) are as follows:

(16) a. $P(A|I,S) = P(r0|START) \times P(r1|NP:fa)) \times P_{lex}(more, NP/N|more') \times P_{lex}(doggies, N|dogs') = \frac{1 \times 0.\dot{3} \times 0.\dot{3} \times 0.\dot{3}}{\sum_d P(d|I,S)}$

b. $P(B|I,S) = P(r0|START) \times P(r2|NP:fa)) \times P_{lex}(doggies, NP\backslash N|more') \times P_{lex}(more, N|dogs') = \frac{1 \times 0.\dot{3} \times 0.\dot{3} \times 0.\dot{3}}{\sum_d P(d|I,S)}$

c. $P(C|I,S) = P(r0|START) \times P_{lex}(more\ \ doggies, NP|more') \times P_{lex}(more\ \ doggies, NP|dogs') = \frac{1 \times 0.\dot{3} \times 0.3 \times 0.\dot{3}}{\sum_d P(d|I,S)}$

$P(A|I,S) = P(B|I,S) = P(C|I,S) = 0.\dot{3}$

Thus, the further increment (c) due to posterior expected frequency can be calculated, to determine $fexp(1)$. In the case of this first sentence, $fexp_{posterior} = fexp_{prior}$, so that $fexp_{posterior} - fexp_{prior} = 0$, and the implicit $\kappa = 1$for all rules.

This means that the initial model must be calculated as follows:

(17) *The Child's First Parsing Model*:

Rule	$fexp(n-1)$	$\frac{(n-1)fexp(n-1) + fexp_E}{n}$	$fexp(n)$
r0. $START \rightarrow NP:fa$	0	1.0	1.0
r1. $NP:fa \rightarrow NP/N:f \quad N:a$	0	$0.\dot{3}$	$0.\dot{3}$
r2. $NP:fa \rightarrow N:a \quad NP\backslash N:f$	0	$0.\dot{3}$	$0.\dot{3}$
l1. $NP/N:more' \rightarrow$ more	0	$0.\dot{3}$	$0.\dot{3}$
l2. $NP\backslash N:more' \rightarrow$ doggies	0	$0.\dot{3}$	$0.\dot{3}$
l3. $N:dogs' \rightarrow$ doggies	0	$0.\dot{3}$	$0.\dot{3}$
l4. $N:dogs' \rightarrow$ more	0	$0.\dot{3}$	$0.\dot{3}$
l5. $NP:more'dogs' \rightarrow$ more doggies	0	$0.\dot{3}$	$0.\dot{3}$

Thus, by (3), we have the following updated probabilistic lexicon:

(18) *The Child's First Lexicon*:

ϕ	σ,λ	f_{exp}	$P_{lex}(\sigma,\lambda\|\phi)$	$P_{lex}(\phi\|\lambda)$
more:=	$\mathbf{NP/N} : \mathbf{more'}_{((e,t),e)}$	$0.\dot{3}$	$0.\dot{3}$	$0.\dot{3}$
	$N : dogs'_{(e,t)}$	$0.\dot{3}$	$0.\dot{3}$	$0.\dot{3}$
doggies:=	$NP\backslash N : more'_{((e,t),e)}$	$0.\dot{3}$	$0.\dot{3}$	$0.\dot{3}$
	$\mathbf{N} : \mathbf{dogs'}_{(e,t)}$	$0.\dot{3}$	$0.\dot{3}$	$0.\dot{3}$
more doggies:=	$NP : (more'dogs')_e$	$0.\dot{3}$	$0.\dot{3}$	$0.\dot{3}$

Since the word counts and conditional probabilities for "more" and "doggies" with them meaning $more'_{((e,t),e)}$ are all equal at this stage, the child may well make errors of overgeneration, using some approximation to "doggies" to mean "more". [5] However, even on the basis of this very underspecified lexicon, the child will not overgenerate "*doggies more". [6]

Let us suppose that the second utterance the child hears is "More cookies". There are again three derivations parallel to (8). The child can derive a new parsing model by adding new rules, updating expected frequencies for all rules in the new set of derivations, and recalculating a posteriori expected frequencies as described:

First, on the basis of the intermediate value $\frac{(1)fexp(1)+fexp_{prior}}{2}$,the length-weighted relative conditional probabilities $P(d|I,S)$ of the three derivations for "More cookies" parallel to (8) are as follows:

(19) a. $P(A|I,S) = P(r0|START) \times P(r1|NP : fa)) \times P_{lex}(more,NP/N|more') \times P_{lex}(cookies,N|cookies') = \frac{1.0 \times 0.\dot{3} \times 0.\dot{3} \times 0.1\dot{6}}{\sum_d P(d|I,S)} = \mathbf{0.42}$

b. $P(B|I,S) = P(r0|START) \times P(r2|NP : fa)) \times P_{lex}(cookies,NP\backslash N|more') \times P_{lex}(more,N|cookies') = \frac{1 \times 0.\dot{3} \times 0.1\dot{6} \times 0.1\dot{6}}{\sum_d P(d|I,S)} = \mathbf{0.23}$

c. $P(C|I,S) = P(r0|START) \times P_{lex}(more \quad cookies,NP|more') \times P_{lex}(more \quad cookies,NP|cookies') = \frac{1 \times 0.\dot{3} \times 0.01\dot{6} \times 0.25}{\sum_d P(d|I,S)} = .\mathbf{35}$

$\mathbf{P(A|I,S) \neq P(B|I,S) \neq P(C|I,S) \neq 0.\dot{3}}$

The further increment (c) due to posterior expected frequency can then be calculated, to determine $fexp(2)$. In the case of the second and all subsequent sentences, $fexp_{posterior} \neq fexp_{prior}$, so that $fexp_{posterior} - fexp_{prior} \neq 0$, and the implicit $\kappa < 1$ for all rules.

The model can then be calculated as:

(20) *The Child's Parsing Model #2*:

Rule	$fexp(n-1)$	$\frac{(n-1)fexp(n-1)+fexp_E}{n}$	$fexp(n)$
r0. $START \rightarrow NP : fa$	1.0	1.0	1.0
r1. $NP : fa \rightarrow NP/N : f \quad N : a$	$0.\dot{3}$	$0.\dot{3}$	0.34
r2. $NP : fa \rightarrow N : a \quad NP\backslash N : f$	$0.\dot{3}$	$0.\dot{3}$	0.25
l1. $NP/N : more' \rightarrow$ more	$0.\dot{3}$	$0.\dot{3}$	0.34
l2. $NP\backslash N : more' \rightarrow$ doggies	$0.\dot{3}$	$0.1\dot{6}$	$0.1\dot{6}$
l3. $N : dogs' \rightarrow$ doggies	$0.\dot{3}$	$0.1\dot{6}$	$0.1\dot{6}$
l4. $N : dogs' \rightarrow$ more	$0.\dot{3}$	$0.1\dot{6}$	$0.1\dot{6}$
l5. $NP : more'dogs') \rightarrow$ more doggies	$0.\dot{1}$	$0.1\dot{6}$	$0.1\dot{6}$
l6. $NP : more'cookies' \rightarrow$ more cookies	0	$0.1\dot{6}$	0.17
l7. $NP\backslash N : more' \rightarrow$ cookies	0	$0.1\dot{6}$	0.11
r8. N(cookies) : $cookies' \rightarrow$ cookies	0	$0.1\dot{6}$	0.24
l9. N(more) : $cookies' \rightarrow$ more	0	$0.1\dot{6}$	0.11

Thus, by (3), we have the following updated probabilistic lexicon:

[5] The example is based on an attested case of this particular overgeneralization (Cathy Urwin, p.c., c. 1979).

[6] It follows that overgeneralizations by the child like "Allgone doggies" must arise from processes of lexical generalization of the category for "more" to a meaning *allgone'* of the same semantic type as *more'*.

(21) *The Child's Lexicon #2*:

ϕ	σ,λ	$exp_{lex}(n)$	$P(\sigma,\lambda\|\phi)$	$P(\phi\|\sigma,\lambda)$
more:=	$\mathbf{NP/N} : \mathbf{more}'_{((e,t),e)}$	$\mathbf{0.34}$	$\mathbf{0.57895}$	$\mathbf{0.57895}$
	$N : dogs'_{(e,t)}$	$0.16\dot{6}$	0.26318	0.5
	$N : cookies'_{(e,t)}$	0.11	0.15789	$0.3\dot{3}$
doggies:=	$NP\backslash N : more'_{((e,t),e)}$	$0.16\dot{6}$	0.5	0.385
	$\mathbf{N} : \mathbf{dogs}'_{(e,t)}$	$0.16\dot{6}$	0.5	0.50
cookies:=	$NP\backslash N : more'_{((e,t),e)}$	0.11	$0.3\dot{3}$	0.15789
	$\mathbf{N} : \mathbf{cookies}'_{(e,t)}$	$\mathbf{0.24}$	$\mathbf{0.6\dot{6}}$	$\mathbf{0.6\dot{6}}$
more doggies:=	$NP : (more'dogs')_e$	$0.16\dot{6}$	$0.3\dot{3}$	$0.3\dot{3}$
more cookies:=	$NP : (more'cookies')_e$	0.17	$0.3\dot{3}$	$0.3\dot{3}$

It should be remarked that the expected frequencies in this table are not quite the same as those that would be obtained by recomputing f_{exp} over the entire corpus, as in standard batch EM. If we did that, then we would realize that the expected frequency of more:= $N : dogs'$ is actually less than the value shown, and that of more := $NP/N : more'$, greater, among other differences. However, this approximation will become more exact as more sentences are analyzed, and it is worth tolerating it in order not to have to make the psychologically implausible and computationally intractable assumption that the child keeps a corpus of all analyses of all sentences it has ever encountered and recomputes the entire model from scratch at each iteration.

In particular, despite the inexactness of the lexical expected frequencies, the probability that the child will correctly say "more" when they mean *more*′ is already greater than that of spurious candidates like "doggies" or "cookies.

Indeed, as this process continues with utterances like "bad doggies" and "bad cookies", the child is exponentially less likely to generate "doggie" when she means "more". The reader should be able to satisfy themselves that this effect will be even stronger for more realistic corpora in which the frequency distribution of words is highly skewed, with open class words like "doggie" being exponentially rarer (hence with lower values for $P(\phi)$) than closed class words like "more". Experimental sampling by elicitation of child utterances during such exponential extinction may well give the appearance of all-or-none lexical learning and setting of parameters like NEG-placement and *pro*-drop claimed by Thornton & Tesan (2006).[7]

This lexicon includes non-standard holophrastic lexical items such as "more doggies". Such spurious lexical entries can later be pruned if necessary on grounds of low relative expected frequency in the corpus as a whole, along with the spurious entries. Nevertheless, holophrastic lexical items such as "All gone," may be sufficiently common as to be useful in their own right, and persist in the developing lexicon in parallel with their components.

It is of course possible that the adult will on occasion mistake the proposition that the child has in mind, or that the child will choose such a proposition wrongly, leading to false lexical associations. However, provided the two get it right most of the time, the same process of Bayesian re-estimation of conditional probabilities of these lexical hypotheses for each word will allow the latter to arrive at a correct lexicon.

Unrevised from here

4.3 Non-Uniform Priors

Until now, we have assumed that the prior conditional probabilities of the lexical syntactic types given the semantic types are uniformly distributed. However, let us suppose that $P(\sigma|\tau)$ is *not* uniform, but rather favors NP/N over $NP\backslash N$ as the type of determiners of type $((e,t),t)$.

On the basis of the intermediate value $\frac{(0)exp(0)+exp_{prior}}{1}$, the relative conditional probabilities $p(d|i,s)$ of the three derivations (8) are now as follows:

[7]This effect is related to the "winner-take-all" effect observed in Steels' 2004 game-based account of the otherwise rather different process of establishing a shared vocabulary among agents who have no preexisting language.

(22) a. $P(A|I,S) = P(r0|START) \times P(r1|NP:fa)) \times P_{lex}(more, NP/N|more') \times P_{lex}(doggies, N|dogs') = \frac{1 \times 0.5 \times 0.\dot{3} \times 0.\dot{3}}{\sum_d P(d|I,S)} = \mathbf{0.5}$

b. $P(B|I,S) = P(r0|START) \times P(r2|NP:fa)) \times P_{lex}(doggies, NP\backslash N|more') \times P_{lex}(more, N|dogs') = \frac{1 \times 0.1\dot{6} \times 0.\dot{3} \times 0.\dot{3}}{\sum_d P(d|I,S)} = \mathbf{0.1\dot{6}}$

c. $P(C|I,S) = P(r0|START) \times P_{lex}(more\ \ doggies, NP|more') \times P_{lex}(more\ \ doggies, NP|dogs') = \frac{1 \times 0.\dot{3} \times 0.3 \times 0.\dot{3}}{\sum_d P(d|I,S)} = \mathbf{0.\dot{3}}$

$P(A|I,S) \neq P(B|I,S) \neq P(C|I,S)$

Thus, the further increment (c) due to posterior expected frequency can be calculated as before, to determine $fexp(1)$. As with the uniformly distibuted prior, in the case of this first sentence, $fexp_{posterior} = fexp_{prior}$, so that $fexp_{posterior} - fexp_{prior} = 0$, and the implicit initial κ is 1 for all rules.

This means that the initial model must be calculated as follows:

(23) *The Child's Parsing Model #1′*:

Rule	$fexp(n-1)$	$\frac{(n-1)fexp(n-1)+fexp_E}{n}$	$fexp(n)$
r0. $START \rightarrow NP:fa$	0	1.0	1.0
r1. $NP:fa \rightarrow NP/N:f \quad N:a$	0	0.5	0.5
r2. $NP:fa \rightarrow N:a \quad NP\backslash N:f$	0	$0.1\dot{6}$	$0.1\dot{6}$
l1. $NP/N:more' \rightarrow$ more	0	0.5	0.5
l2. $NP\backslash N:more' \rightarrow$ doggies	0	$0.1\dot{6}$	$0.1\dot{6}$
l3. $N:dogs' \rightarrow$ doggies	0	0.5	0.5
l4. $N:dogs' \rightarrow$ more	0	$0.1\dot{6}$	$0.1\dot{6}$
l5. $NP:more'dogs' \rightarrow$ more doggies	0	$0.\dot{3}$	$0.\dot{3}$

Thus, by (3), we have the following updated probabilistic lexicon:

(24) *The Child's Lexicon #1′*:

ϕ	σ,λ	f_{exp}	$P_{lex}(\sigma,\lambda\|\phi)$	$P_{lex}(\phi\|\lambda)$
more:=	$\mathbf{NP/N}:\mathbf{more}'_{((e,t),e)}$	0.5	0.75	0.75
	$N:dogs'_{(e,t)}$	$0.1\dot{6}$	0.25	0.25
doggies:=	$NP\backslash N:more'_{((e,t),e)}$	$0.1\dot{6}$	0.25	0.25
	$\mathbf{N}:\mathbf{dogs}'_{(e,t)}$	$0.5\dot{3}$	0.75	0.75
more doggies:=	$NP:(more'dogs')_e$	$0.\dot{3}$	$0.\dot{3}$	$0.\dot{3}$

Since the word counts and conditional probabilities for "more" and "doggies" with them meaning $more'_{((e,t),e)}$ are no longer equal at this stage, the child is less likely to make errors of overgeneration in English, using "doggies" to mean "more". This observation also implies that if the target language *is* in fact determiner-final, then the child is *more* likely to make an error after exposure to a similar first sentence. However, since positive evidence against this bias from determiner-final equivalents of "More cookies" etc. will subsequently become available, the bias will be ovedrridden, early in the acquisition process, for exactly the same reason that, in the absence of such biases, the partially developed lexicon (21) defines a learned distribution $P(\sigma|\tau)$ which will bias the child towards the categories of English.

5 Transitives and Semantic and pragmatic Innate Priors

Both kinds of bias, innate and learned, play a part in the learning of transitive verbs in English and other languages, and show that innate priors may be semantically and pragmatically determined, rather than via some syntactically specific "language instinct".

Unlike intransitive predicates and the determiner category considered in section 4.2, transitive verbs as presented in examples like the following could in principle be assigned either of the two syntactic categories in (26), both of which support a derivation of the logical form:[8]

(25) I see you! := $S:see'you'i'$

[8] We continue to assume for the sake of simple exposition that there is only one logical form supported by the context. In particular, we assume that the corresponding passive is not salient, or that if it is it has a distinct logical form from the active. We will abandon these restrictions later.

(26) a. see := $(S\backslash NP)/NP : \lambda x\lambda y.see'xy$
b. see := *$(S/NP)\backslash NP : \lambda y\lambda x.see'xy$

No SVO language/construction has ever been seriously argued to have a surface syntax corresponding to the second category. We can therefore safely assume that it is either not included at allin the universal set of possible syntactic categories for interpretations of type $(e,(e,t))$, or has an extremely low prior probability.

Specifically, we will assume that the universally permitted set of transitive categories is the following, corresponding to the six basic constituent orders, here listed in order of decreasing frequency of attestation of the order in question.[9]

(27) a. SOV := $(S\backslash NP)\backslash NP : \lambda x\lambda y.see'xy$
b. SVO := $(S\backslash NP)/NP : \lambda x\lambda y.see'xy$
c. VSO := $(S/NP)/NP : \lambda y\lambda x.see'xy$
d. VOS := $(S/NP)/NP : \lambda x\lambda y.see'xy$
e. OVS := $(S/NP)\backslash NP : \lambda x\lambda y.see'xy$
f. OSV := $(S\backslash NP)\backslash NP : \lambda y\lambda x.see'xy$

The decreasing frequency of these orders appears to reflect two independent defeasible constraints. One favors linearization of subject before object. The other favors keeping the syntactic command relations between subject and object as reflected in order of combination the same as those in the logical form.[10]

The first of these constraints appears to be information-structural or "functional" in the sense of the Prague School. The second constraint appears to concern reducing complexity in the syntax-semantics interface. Thus. neither is specifically syntactic, and in fact either may ultimately derive from the nature of animal interaction with the socio-pragmatic world. Thus, the term "innate" is used in a very weak sense here, to mean "not learned by individuals", rather than in the strong sense of "specifically evolved" or "genetically programmed language instinct".

Since (26b) violates the second of these constraints, we are justified in assuming it has a lower prior. Thus the child faced with the pair (25) effectively has only one candidate category for the transitive verb. However, this does not exhaust the problem of learning transitive verbs, because a context may support more than one of these categories.

5.1 Contextual Ambiguity and Learned Priors

Many languages, perhaps all, allow a number of lexical alternations of transitives, as in the case of English "chase/flee" where the same physical situation seems to support more than one logical form. How do children faced with (artificial) examples like the following avoid the error of making an OVS lexical entry for "flee" with the meaning *chase*′?

(28) Kitties flee doggies!

It is important that examples of the verb class of which "flee" is the most common representative are rare. In particular, in comparison to 162 occurences of inflected forms of the verb "chase," there is exactly one occurrence of any form of "flee" in the entire CHILDES corpus. We are therefore justified in assuming that the child will have encountered plenty of unambiguous transitive verbs in utterances like (25) before encountering examples like (28).

This means that the learned prior probability of the instantiated rules for combining transitive SVO verbs with their object and subject will be substantially greater than the priors for OVS verbs by the time they eventually do encounter (28), for much the same reason that the probability of the rule r1 is greater than that of r2 in (21). For the sake of illustration let's conservatively assume they have seen 1000 tokens—and adds one count each for these two categories. In that case, since $P(\lambda_\tau|\phi)$ is the same for both, and $P((S\backslash NP)/NP|$"flee") is $\frac{.25\cdot 1000}{1001} = .25$, while $P((S/NP)\backslash NP|$"chase") is $\frac{.25\cdot 1}{1001} = .00025$, the lexical probability for the two entries stand in a ratio of 1000:1. Others are that "flee" means *cats*′, that "kitties" means *flee*′, etc.. However, on the assumption that the child has previously encountered the words "kitties" and "doggies" with reasonably high relative frequency, so that "flee" is the only unfamiliar word in the sentence, the probability model for derivations will, by definition (3), assign a very low expected frequency, and hence low probability, to these spurious hypotheses.

[9] We assume, following Baldridge (2002), that free word-order languages simply have more than one of these categories.

[10] Two of these categories, VSO and OSV, "wrap" their most oblique argument O(bject) around their least oblique argument S(ubject). (These categories are forced under the account of argument cluster coordination and the restriction to the combinators **BTS** in CCG—Steedman 2000b).

Thus, *provided that the adult's intended meaning is available*, even if with low prior probability, then the child is in a position to assign the correct hypothesis a high probability. (Even if it is not available, the child will assign a low probability to the spurious lexical entry for *chase′*.)

Gleitman 1990; Gleitman et al. 2005 has described the process by which the child resolves contextual ambiguity as "syntactic bootstrapping," meaning that it is the childs knowledge of the language-specific grammar, as opposed to the semantics, that guides lexical acquisition. However, in present terms such an influence on learning is simply emergent from the statistical model used in semantic bootstrapping. We will return to this point in the Conclusion.

5.2 Against Parameters

Like the related proposals of Siskind; Villavicencio; Zettlemoyer et al. and the somewhat different probabilistic approach of Yang 2002, this proposal considerably simplifies the logical problem of language acquisition. In particular, it allows us to eliminate the Subset Principle of Berwick (1985), and attendant requirements for ordered presentation of unambiguous parametric triggers, both of which appear to present serious problems for the language learner (Angluin 1980; Becker 2005; Fodor & Sakas 2005). Nor does this move contradict widely-held assumptions concerning the "poverty of the stimulus", and in particular the unavailability to the child of negative evidence. The child's progression from the universal superset grammar to the language-specific target grammar is entirely determined by positive evidence raising the probability of correct hypotheses at the expense of incorrect ones. The incorrect hypotheses that are eliminated in this way include any that are introduced by error and noise. The only evidence that the child needs in order to learn their language is a reasonable proportion of utterances involving sentences which are sufficiently short for them to deal with.

The theory presented here resembles the proposal of Fodor 1998 as developed in Sakas & Fodor (2001) and Niyogi (2006) in that it treats the acquisition of grammar as arising from parsing with a universal "supergrammar". As in that proposal, both parameters and triggers are simply properties of the language-specific grammar itself—in their case, rules over independently learned parts of speech, in present terms, lexical categories.

It differs in assuming that the unordered logical form for the utterance is mostly available, with tolerable degrees of error and ambiguity. This means that the problem of syntactically ambiguous sentences to which STL is heir does not arise.

It also differs in the algorithm by which it converges on the target grammar. Rather than learning rules in an all or none fashion on the basis of unambiguous sentences that admit of only one analysis, it adjusts probabilities in a model of all elements of the grammar for which there is positive evidence for *all* processable utterances. In this respect, it more closely resembles the proposal of Yang (2002). However it differs from both in eschewing the view that grammar learning is parameter setting.

If the parameters are implicit in the rules or categories themselves, and you can learn the rules or categories directly, why should the child or the theory bother with parameters at all? For the child, all-or-none parameter-setting is counterproductive, as it will make it hard to learn the many languages which have inconsistent settings of parameters across lexical types and exceptional lexical items, as in German and Dutch head finality, and English expressions like the following:

(29) Doggies galore!

Therefore, the fact that languages show violable tendencies to consistency for values of parameters like headedness across categories for related semantic types such as verbs and prepositions probably stems from considerations of overall encoding efficiency for the grammar as a whole, of the kind captured in notions like Minimal Description Length (MDL). Such considerations may be relevant to comparing entire grammars for the purpose of explaining language change, as in the work of Briscoe (2000). Their presence will under the present theory make the task of learning easier, by raising prior probabilities in the model for rules and categories that actually do recur. But it is less clear that representing them explicitly, rather than leaving them implicit in th model, will help the individual child learning a specific grammar, word-by-word.

6 A More Realistic Lexicon

If children's exposure to language were merely confined to recitations of propositions they already had in mind, it would be a dull affair. It is not even clear why they would bother to learn language at all, as Clark

(2004) points out in defence of a PAC learning model.[11]

However, the worked example above is deliberately simplified in respect of the child's syntax and semantics. We know from Fernald et al. (1989) and Fernald (1993) that infants are sensitive to interpersonal meanings of intonation from a very early age. In English, intonation contour is used to convey a complex system of information-structural elements, including topic/comment markers and given/newness markers (Bolinger 1965; Halliday 1967; Ladd 1996), and is exuberantly used in speech by and to infants. It is this part of the meaning that constitutes the whole point of the exercise for the child, providing the motivation that Clark questions.

For example, it is likely that the child's representation of the utterance "MORE DOGGIES! is more like (30), which uses the notation of Steedman 2000a, 2007, in which [S] represents speaker modality (contributed by the LL% boundary tone), ρ

(30)
$$\frac{\dfrac{\text{MORE DOGGIES}}{\underset{NP_{+,\rho}{}^{\uparrow}}{\text{H}*\quad \text{H}*}}}{\underset{NP^{\uparrow}{}_{\phi} : [S]\rho\lambda p.p(*more'*dogs')}{:\lambda p.p(*more'*dogs')}} \quad \frac{\dfrac{!}{\text{LL\%}}}{\underset{:\lambda g.\pi[\text{S}]\eta\ g}{X_{\phi}\backslash_{*}X_{\pi,\eta}}} \;<$$

"Mummy makes the property afforded by more dogs common ground."

The semantics of speech acts goes beyond the immediate concerns of this paper, and is discussed by Steedman & Petrick (2007), who note that the inference system that the semantics supports is cosely rrelated to that involved in planning with sensing actions.

The set of type-raised NP categories licenced by UG that is schematized in (30) as $NP^{\uparrow}$ denotes the set of all order-preserving functions over functions-over-NP onto the results of applying those functions to the original NP. It includes categories of the following two forms, where T is a variable over all category types::

(31) $\text{T}/(\text{T}\backslash\text{NP}) : \lambda\text{p}\lambda\text{x.px}$
$\text{T}\backslash(\text{T}/NP) : \lambda p\lambda x.px$

We also need the following related non-order-preserving "extracting" categories, in which S_x indicates a distinct type of clause:

(32) $S_x\backslash(S\backslash NP) : \lambda p\lambda x.px$
$S_x/(S/NP) : \lambda p\lambda x.px$

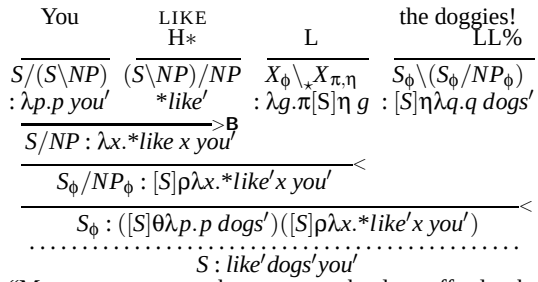
While, up until now, we have only seen one syntactic type per semantic type in the child's lexicon for English, in general a single semantic type may be realized by many syntactic types in a single language, and this is the case for English NPs. Such ambiguity is perfectly compatible with the learning procedure defined earlier: it just means that there will be several categories with substantial conditional probability mass $P(\sigma|\phi)$

It may seem surprising that a language should allow so much ambiguity in such a basic linguistic category type as NP. However, this is simply the same proliferation of syntactic types that would be disambiguated in a language with overt morphological *case*. English just happens to be a language which has so-called structural case, implicit in linear order. We shall see that the child will not find this a problem. But first we need to consider the role of intonation in the child's grammar.

While intonation has been shown to be even more markedly discrepant from traditional syntactic structure in child-directed and child-originated speech (Fisher & Tokura 1996; Gerken et al. 1994; Gerken 1996) than in adult dialog, in CCG intonation structure is united with a freer notion of derivational structure. Consider the child in a similar situation faced with the following utterance, from Fisher & Tokura (1996) as discussed in Steedman 1996:

[11] On the other hand, it is equally unclear why they would bother to learn language in the *absence* of any interpretation, as assumed in PAC learning, a point which Clark doesn't address.

(33)



"Mummy supposes what property the dogs afford to be common ground, Mummy makes it common ground it's me liking them."

Fisher points out that the L intermediate phrase boundary that she observed after the verb makes the intonation structure inconsistent with standard assumptions about surface constituency. However, this intonation structure is isomorphic to the CCG derivation above, which delivers the corresponding theme/rheme information partition directly.

Thus, here too, the availability of the full semantic interpretation, including information-structural information, directly reveals the target grammar. In this case, since the derivation requires the use of the forward composition rule, indexed $>\mathbf{B}$, the child gets information not only about the probability of the verb, the nominative, and the accusative categories of English, but also about the probability of applying the composition rule to the first two categories, the probability that the subject of "like" will be headed by "you", and its object be headed by "doggies". Thus, the child can build the parsing model in parallel with learning the grammar.

Zettlemoyer et al. 2005 do not include a parser model as distinct from the lexical model, and they get away with this because their grammar is small. For small unambiguous grammars, CCG categories alone are usually enough to eliminate search. However, we know from experience with parsers of the size of those needed for the Wall Street Journal corpus that such a model will be necessary once the child's grammar begins to approach adult size.

It is not clear whether the child uses a head-dependency model of the kind used by Collins (1999) as a parsing oracle, or uses semantics and world knowledge directly, as proposed by Crain & Steedman 1985 and others, or some interpolation of the two. Some further details of the former kind of model and one possible algorithm by which it can be learned incrementally are set out in the appendix.

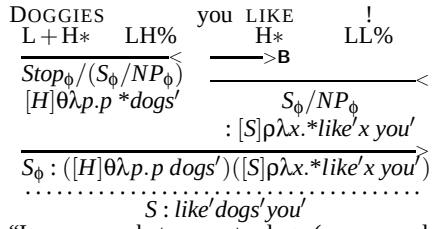
7 SMOOTHING AND GENERALIZATION

A standard assumption in wide-coverage parsing using treebank grammars is that the grammar must be generalized and the statistical model must be smoothed with respect to unseen words and word-category pairs. Since all language-specific information in CCG resides in the lexicon, this amounts to predicting unseen word-category pairs and head-dependencies.

Generalizing grammars is a tricky business: Fodor & Sakas offer as an example the observation that the child should assume on the basis of seen topicalizations in English that all NPs can undergo topicalization. However, they should not assume on the basis of observations of negative placement with repect to auxilliaries that the same process can apply to all verbs.

This problem looks rather different from the present perspective. Since we are learning a probabilistic instance of universal grammar, the grammar is already generalized, and predicts all possible word-category pairs. Since topicalization is a lexically-specified contruction in CCG, when the child hears the following as its first example of the construction, it still has available all possible categories for "doggies", including the preposing topicalized one that supports this derivation:

(34)



"I suppose what property dogs (as opposed to something else) afford to be common ground, Mummy makes it common ground it's me liking them."

So the conditional probability of this category given this type $P(Stop_\phi/(S_\phi/NP_\phi)|((e,t),t)$ will grow and become available to other words, supporting generalization.

We must correspondingly assume that the non-generalization of the negative category is based on a semantically distinct type of verb.

8 Grounding Semantics in Interaction

One might ask at this point how the child or machine comes to have access to the logical form $more'dogs'$ (or whatever), and why she does not entertain other candidates, such as $more'tails'$. As Quine (1960) pointed out, this is a different kind of question, whose answer lies in the nature of the child's sensory-motor interactions with the world, and depends as much on mammalian evolution as on learning in the individual child.

Nevertheless, this observation carries a warning that the semantics that emerges from that interaction and those evolutionary processes may be very unlike the semantics that naive logicist assumptions suggest, and that is found in the corpora of database queries used by Thompson & Mooney and Zettlemoyer et al.. For example, the logical form that the child brings to (33) is likely be something more like $give'pleasure'you'dogs'$, so that the lexical entry for "like" of type $(e,(e,t))$ is the following, exhibiting the same "quirky" relation between (structural) nominative case and an underlying dative role that Icelandic exhibits morphologically for the corresponding verb:

(35) like := $(S\backslash NP)/NP : \lambda x\lambda y.give'pleasure'\ y\ x'$

Similarly, it is quite possible that the childs initial representation of the meaning of "more" is as a predicate $S/NP : more$, and that it is the resulting prior on the conditional probability $P(S/NP|e \rightarrow t)$ that is generalized to "allgone", leading to transient non-standard orders like "Allgone milk". Or "all gone" may be misanalysed as a proto-determiner like "no more." These questions are much harder to investigate. While one can annotate corpora such as CHILDES with logical forms, as Villavicencio did, one has very little idea of what relation such logical forms bear to a psychologically real adult semantics, let alone a child's. This fact makes quantitative testing of the present theory difficult.

One we way around this is to do linguistics, meditating on the huge collection of phenomena to do with binding, case, classification, tense and aspect, and so on, that seem to dimly reveal an underlying system of meanings, in the hope of discerning the real semantics. This is a very hard problem, and progress seems slow.

Another alternative is to investigate the question qualitatively, using simulated language learners. Since large corpora of artificial logical forms such as database queries annotated with sentences are unlikely to become available, and everyone believes that the semantics is determined by the child's sensory-motor experience of acting in the physical world, this makes the use of physically grounded robots particularly interesting. Projects of this kind are under investigation by a number of groups, including those led by Luc Steels, Deb Roy, and Geert-Jan Kruijff. These groups are looking at emergence of agreed vocabulary among prelinguistic agents (Steels & Baillie 2003; Steels 2004), plans and plan-recognition as a basis for situated language understanding (Roy 2005; Gorniak & Roy 2007), and context-dependent spatial models for natural language semantics (Kelleher et al. 2006). However, these projects so far rely on forms of semantics that are designed top-down, using the robot tasks as a forcing function, rather than on a semantics developed bottom-up from action representations themselves. Delivering semantic representations that are grounded in the same sense that mechanisms developed over hundreds of millions of years of evolution is much harder. Steedman (2002b,a) argues that the combinators **B** and **T** that do most of the projective syntactic work in CCG are directly related to operations of seriation and affordance in the planner. This suggests that mechanisms for state-based reactive planning of the kind investigated by Petrick & Bacchus (2002, 2004) may offer a way towards a more distinctively action-based semantics for natural language (cf. Steedman 2008, Geib & Steedman 2007).

9 Conclusion

This paper has argued that syntax is learned on the basis of preexisting semantic interpretations afforded by the situation of adult utterance, using a generative statistical model over a universal set of grammatical possibilities. The existence of the model itself helps the child to rapidly acquire a correct grammar even in the face of competing ambiguous semantics.

In equating language-specific grammar with a statistical model for parsing with universal grammar, the proposal bears an intriguing relation to the Maximum Spanning Tree (MST) parser (McDonald et al. 2005; McDonald & Pereira 2006b,a). This parser searches for the maximum-valued spanning tree-forming subgraph of a totally connected graph over the words of the string, using a perceptron-like maximum-margin discriminative model trained using pairs of strings and dependency trees. It has been applied to parsing "non-projective" or long-range dependencies, including crossing dependencies. It works best when the features over which the model is trained are grammar-like features such as position with respect to the verb, or morphological features. In particular, Çakıcı (2007) has shown that using CCG categories as features in a dependency-model of Turkish improves performance over the baseline in McDonald & Pereira (2006b). MST could therefore be seen as offering an alternative, discriminative, version of the present approach, according to which it could be used to learn weights for a language-specific set of features or categories drawn from a larger universal set.

The fact that the onset of syntactically productive language at the end of the Piagetian sensory-motor develomental phase is accompanied by an explosion of advances in qualitatively different "operational" cognitive abilities suggests that the availability of language has a feedback effect, facilitating access to concepts that the child would not otherwise have access. Early work by Oléron (1953) and Furth (1961) on specific cognitive deficits concerning non-perceptually evident concepts arising in deaf children who had been linguistically deprived by being denied access to sign supports this view.

This means that Gleitman's (1990) influential suggestion that it is the availability of syntax that enables the child to "syntactically bootstrap" lexical entries for verbs (such as "think") that are not situationally evident is essentially correct. However, we have seen from the case of learning the verb "flee" in the face of competition from the meaning *chase′* that it is the availability to the child of *a model of the relation between language-specific syntax and universal semantics* that makes this possible. It follows that the effects observed by Oléron and Furth, and Gleitman herself must have the character of *directing the child's attention* to alternatives that are available to them, but which they would otherwise overlook, by sheer force of Bayesian priors on the conditional probability $P(\sigma|\tau)$ of a syntactic category given a semantic type, as seems to be implicit in Gleitman et al. (2005). In that sense, we should probably refer to this effect as "grammatical" boostrapping, since it is an effect that is both syntactic and semantic.

ACKNOWLEDGEMENTS

The work was supported in part by the SE Edinburgh-Stanford Link grant and EU IST grant FP6-2004-IST-4-27657 PACO-PLUS. Thanks to Cathy Urwin for the example and other insights into child language acquisition.

APPENDIX A: PARSING MODELS

A *grammar G* defines a language L as a set T of trees, such that the yield of each tree $t \in T$ is a string $s \in L$. A *probabilistic grammar* defines a probability distribution over these trees, i.e.:

(36) a. $\forall t \in T[0 \leq P(t) \leq 1]$
b. $\sum_{t \in T} P(t) = 1$

We can define the most likely tree *tree* for a given sentence *sentence* as:

$$\text{(37)} \qquad \arg\max P(tree|sentence) = \underset{tree \in parses(sentence)}{\arg\max} \frac{P(tree, sentence)}{P(sentence)} = \underset{tree \in parses(sentence)}{\arg\max} P(tree)$$

where

$$\text{(38)} \quad P(tree) = \prod_{production \in tree} P(production_i | parentCategory_i)$$

That is, $P(tree)$ is computed as the product of the probabilities of all its productions, as in Figure 1, where the individual probabilities are computed as relative frequencies of occurence in a treebank conditional on the parent category of the production:

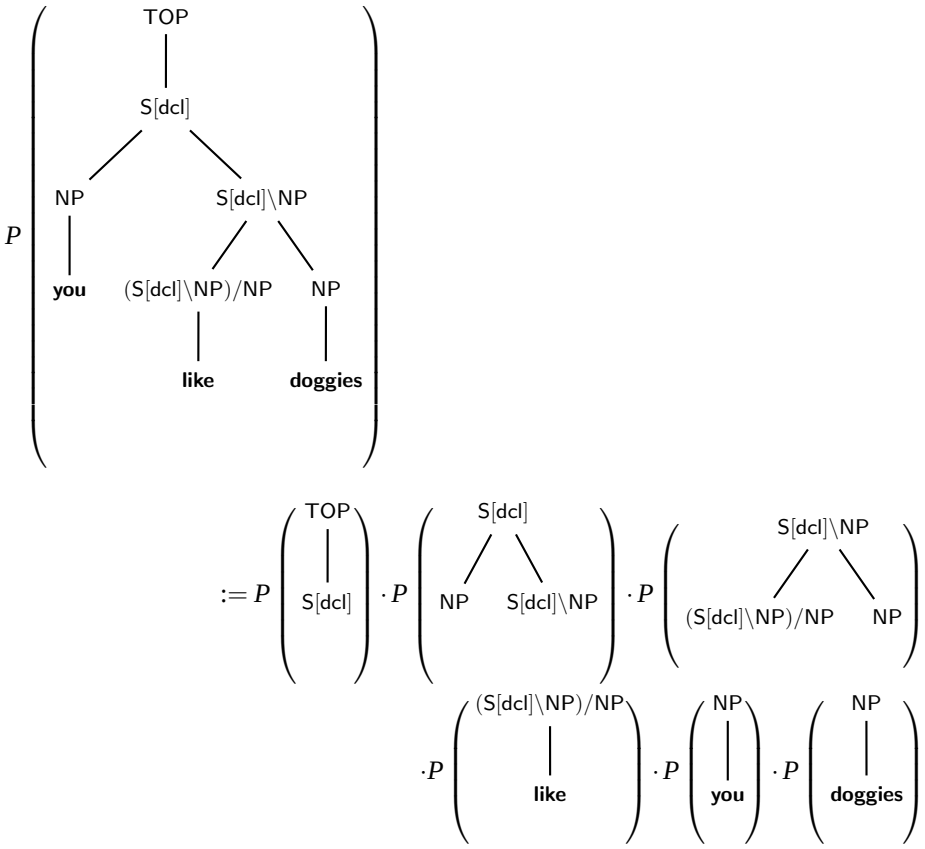


Figure 1: Computing $P(tree)$ for a PCFG

Such a probabilistic grammar is unsound, because it assumes that probabilities such as that of an NP realized as "you" are independent of their role in a larger tree as subject *vs.* object of "like". However, we can generalize the probabilistic grammar to a **head-word dependency model**, in which the probability of such trees is computed as the product of probabilities of productions. Thus $P(tree)$ is given by the product related to (38) in Figure 2, where it is defined algorithmically rather than graphically for reasons of space (we assume **lexicalHead(X) =** ⟨ **headCat(X), headWord(X)** ⟩).

For a simple PCFG in which NT and T symbols are disjoint, we can assume that the probability of a tree, $P(\tau)$ is recursively decomposed into the rule probabilities $P(X \rightarrow \alpha) = P(X \rightarrow \alpha | X)$ (such that

1. Generate TOP and its lexical head

$$P(headCat(TOP) = (\mathsf{S[dcl]}\backslash\mathsf{NP})/\mathsf{NP})$$
$$\cdot P(headWord(TOP) = \textbf{like} \quad | \quad lexCat(TOP) = (\mathsf{S[dcl]}\backslash\mathsf{NP})/\mathsf{NP})$$

2. Generate expansion of TOP

$$\cdot P(tree = \begin{array}{c}\mathsf{TOP}_{parent}\\ |\\ \mathsf{S[dcl]}_{head}\end{array} \quad | \quad parent = \mathsf{TOP},$$
$$lexicalHead(parent) = \langle(\mathsf{S[dcl]}\backslash\mathsf{NP})/\mathsf{NP}, \textbf{like}\rangle)$$

3. Generate expansion of $\mathsf{S[dcl]}$

$$\cdot P(tree = \begin{array}{ccc}&\mathsf{S[dcl]}_{parent}&\\ \mathsf{NP}_{sister}&&\mathsf{S[dcl]}\backslash\mathsf{NP}_{head}\end{array} \quad | \quad parent = \mathsf{S[dcl]},$$
$$lexicalHead(parent) = \langle(\mathsf{S[dcl]}\backslash\mathsf{NP})/\mathsf{NP}, \textbf{like}\rangle)$$

4. Generate lexical head of sister node NP

$$\cdot P(headCat(sister) = \mathsf{NP} \quad | \quad category(sister) = \mathsf{NP})$$
$$\cdot P(headWord(sister) = \textbf{You} \quad | \quad headCat(sister) = \mathsf{NP},$$
$$tree = \begin{array}{ccc}&\mathsf{S[dcl]}_{parent}&\\ \mathsf{NP}_{sister}&&\mathsf{S[dcl]}\backslash\mathsf{NP}_{head}\end{array},$$
$$lexicalHead(parent) = \langle(\mathsf{S[dcl]}\backslash\mathsf{NP})/\mathsf{NP}, \textbf{like}\rangle)$$

5. Generate expansion of $\mathsf{S[dcl]}\backslash\mathsf{NP}$

$$\cdot P(tree = \begin{array}{ccc}&\mathsf{S[dcl]}\backslash\mathsf{NP}_{parent}&\\ (\mathsf{S[dcl]}\backslash\mathsf{NP})/\mathsf{NP}_{head}&&\mathsf{NP}_{sister}\end{array} \quad | \quad parent = \mathsf{S[dcl]},$$
$$lexicalHead(parent) = \langle(\mathsf{S[dcl]}\backslash\mathsf{NP})/\mathsf{NP}, \textbf{like}\rangle)$$

6. Generate lexical head of NP

$$\cdot P(headCat(sister) = \mathsf{NP} \quad | \quad category(sister) = \mathsf{NP})$$
$$\cdot P(headWord(sister) = \textbf{doggies} \quad | \quad headCat(sister) = \mathsf{NP},$$
$$tree = \begin{array}{ccc}&\mathsf{S[dcl]}\backslash\mathsf{NP}_{parent}&\\ (\mathsf{S[dcl]}\backslash\mathsf{NP})/\mathsf{NP}_{head}&&\mathsf{NP}_{sister}\end{array},$$
$$lexicalHead(parent) = \langle(\mathsf{S[dcl]}\backslash\mathsf{NP})/\mathsf{NP}, \textbf{like}\rangle)$$

Figure 2: Computing $P(tree)$ for a Head-Dependency Model

$\sum_i P(X \rightarrow \alpha_i|X) = 1$ and lexical probabilities $P(w|c)$ (such that $\sum_i P(w_i|c) = 1$. We can then estimate the model as follows.

Incremental parameter estimation of a PCFG

We assume that the model is stored as a vector $\mathbf{f}$ of frequency counts, and that

$$P(X \rightarrow \alpha_i|X) = \frac{\mathbf{f}(X \rightarrow \alpha_i)}{\sum_j \mathbf{f}(X \rightarrow \alpha_j)}$$

Learning is an incremental variant of the inside-outside (EM) algorithm for PCFGs (Pereira & Schabes 1992), and proceeds as follows:

1. **Initialize model**

- We assume that the initial model is determined by a frequency vector $\mathbf{f}_0$ which consists of hallucinated (fractional) counts that form a consistent model.
- The magnitude of these counts (e.g. whether they assume we have seen one sentence or a million) effectively determines a learning rate.
- We will assume a count of 1 for every unseen lexical entry or rule that is consistent with some derivation permitted by UG, and zero otherwise.
- We can also think of $\mathbf{f}_0$ as a more interesting innate probability distribution over rules and categories of Universal Grammar.

2. **Parse sentences** Repeat at each time step t:

   (a) **Parse sentence** s_t.
   This yields a set of derivations D_i for s_t, each of which has a probability $P_t(D_i)$, and a conditional probability $P_t(D_i|s_t) = \frac{P_t(D_i)}{\sum_j P_t(D_j)}$, where the denominator sums over all derivations D_j for s_t.
   The expected frequency of a rule $X \rightarrow \alpha_i$ in s_t, $\langle \mathbf{f}(X \rightarrow \alpha_i|s_t)\rangle_{P_t}$ is

   $$\langle \mathbf{f}(X \rightarrow \alpha_i|s_t)\rangle_{P_t} = \sum_j P_t(D_j|s_t) freq(X \rightarrow \alpha_i, D_j)$$

   where $freq(X \rightarrow \alpha, D)$ is the frequency of $X \rightarrow \alpha$ in D. This defines a vector of expected frequencies, $\langle \mathbf{f}(s)\rangle_{P_t}$.

   (b) **Update the probability model**.
   We generate a new vector of (expected) frequency counts, $\mathbf{f}_{t+1}$, by adding $\mathbf{f}(s)\rangle_{P_t}$, the expected frequencies of all rules and word-category pairs in s_t, to $\mathbf{f}_t$. $\mathbf{f}_{t+1} := \mathbf{f}_t + \langle \mathbf{f}(s)\rangle_{P_t}$.

Learning the lexicon

While the set of rules can be assumed to be given by UG, the set of words for a given language certainly isn't. We assume that each lexical category (preterminal) can expand to an "unknown" word *UNK*, and that initially, $P_0(UNK|c) = 1 = \frac{n}{n}$ for all categories (n again determines a learning rate, and should (probably) be 1). Furthermore, we assume that there is a frequency threshold θ, below which the probability of a words is calculcated based on $P_t(UNK|c)$. For each new word, a separate frequency count is kept, and thus

$$P_t(UNK|c) = \frac{\sum_{w:\mathbf{f}_t(w)<\theta} \mathbf{f}_t(w|c) + \mathbf{f}_t(UNK|c)}{\mathbf{f}_t(c)}$$

Note that $\mathbf{f}_t(UNK|c) = \mathbf{f}_0(UNK|c)$, because *UNK* is never actually encountered.

REFERENCES

Angluin, D. (1980). Inductive inference of formal languages from positive data. *Information and Control, 45*, 117–135.

Baldridge, J. (2002). Lexically specified derivational control in combinatory categorial grammar. Ph.D. thesis, University of Edinburgh.

Becker, M. (2005). Raising, control, and the subset principle. In *Proceedings of the 24th West Coast Conference on Formal Linguistics*, pp. 52–60, Somerville MA. Cascadilla Proceedings Project.

Berwick, R. (1985). *The Acquisition of Syntactic Knowledge*. Cambridge, MA: MIT Press.

Bolinger, D. (1965). *Forms of English*. Cambridge, MA: Harvard University Press.

Briscoe, T. (2000). Grammatical acquisition: Inductive bias and coevolution of language and the language acquisition device. *Language, 76*.

Buszkowski, W. & Penn, G. (1990). Categorial grammars determined from linguistic data by unification. *Studia Logica, 49*, 431–454.

Buttery, P. (2006). Computational models for first language acquisition. Ph.D. thesis, University of Cambridge.

Çakıcı, R. (2007). Parser models for a highly inflected language. Ph.D. thesis, University of Edinburgh. in preparation.

Charniak, E. (2000). A maximum-entropy-inspired parser. In *Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics*, pp. 132–139, Seattle, WA.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Chomsky, N. (1995). *The Minimalist Program*. Cambridge, MA: MIT Press.

Clark, A. (2004). Grammatical inference and first language acquisition. In *CoLing Workshop on Psychocomputational Models of Human Language Acquisition*.

Clark, R. & Roberts, I. (1993). A computational model of language learnability and language change. *Linguistic Inquiry, 24*, 299–345.

Collins, M. (1997). Three generative lexicalized models for statistical parsing. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics, Madrid*, pp. 16–23, San Francisco, CA. Morgan Kaufmann.

Collins, M. (1999). Head-driven statistical models for natural language parsing. Ph.D. thesis, University of Pennsylvania.

Collins, M. (2003). Head-driven statistical models for natural language parsing. *Computational Linguistics, 29*, 589–637.

Crain, S. & Steedman, M. (1985). On not being led up the garden path: the use of context by the psychological parser. In L. K. David Dowty & A. Zwicky (eds.), *Natural Language Parsing: Psychological, Computational and Theoretical Perspectives*, pp. 320–358. Cambridge: Cambridge University Press.

Crain, S. & Thornton, R. (1998). *Investigations in Universal Grammar*. Cambridge MA: MIT Press.

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B, 39*, 1–38.

Fernald, A. (1993). Approval and disapproval: Infant responsiveness to vocal affect in familiar and unfamiliar languages. *Child Development, 64*, 657–667.

Fernald, A., Taeschner, T., Dunn, J., Papousek, M., Boysson-Bardies, B., & Fukui, I. (1989). A cross-language study of prosodic modifications in mothers' and fathers' speech to infants. *Journal of Child Language*, *16*, 477–501.

Fisher, C. & Tokura, H. (1996). Prosody in speech to infants: Direct and indirect acoustic cues to syntactic structure. In J. Morgan & K. Demuth (eds.), *Signal to Syntax: Bootstrapping from Speech to Grammar in Early Acquisition*, pp. 343–363. Erlbaum.

Fodor, J. D. (1998). Unambiguous triggers. *Linguistic Inquiry*, *29*, 1–36.

Fodor, J. D. & Sakas, W. (2005). The subset principle in syntax: Costs of compliance. *Journal of Linguistics*, *41*, 513–569.

Furth, H. (1961). The influence of language ofn the development of concept formation in deaf children. *Journal of Abnormal and Social Psychology*, *63*, 386–389.

Geib, C. & Steedman, M. (2007). On natural language processing and plan recognition. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pp. 1612–1617.

Gerken, L. (1996). Prosodic structure in young children's language production. *Langauge*, *72*, 683–712.

Gerken, L., Jusczyk, P., & Mandel, D. (1994). When prosody fails to cue syntactic structure. *Cognition*, *51*, 237–265.

Gibson, E. & Wexler, K. (1995). Triggers. *Linguistic Inquiry*, *25*, 355–407.

Gleitman, L. (1990). The structural sources of verb meanings. *Language Acquisition*, *1*, 1–55.

Gleitman, L., Cassidy, K., Nappa, R., Papafragou, A., & Trueswell, J. C. (2005). Hard words. *Language learning and Development*, *1*, 23–64.

Gorniak, P. & Roy, D. (2007). Situated language understanding as filtering perceived affordances. *Cognitive Science*, *31*. to appear.

Halliday, M. (1967). *Intonation and Grammar in British English*. The Hague: Mouton.

Hockenmaier, J. & Steedman, M. (2002). Generative models for statistical parsing with Combinatory Categorial Grammar. In *Proceedings of the 40th Meeting of the ACL*, pp. 335–342, Philadelphia, PA.

Horning, J. (1969). A study of grammatical inference. Ph.D. thesis, Stanford.

Hyams, N. (1986). *Language Acquisition and the Theory of Parameters*. Dordrecht: Reidel.

Kanazawa, M. (1998). *Learnable Classes of Categorial Grammars*. Stanford, CA: CSLI/folli.

Kelleher, J. D., Kruijff, G.-J. M., & Costello, F. J. (2006). Proximity in context: An empirically grounded computational model of proximity for processing topological spatial expressions. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pp. 745–752, Sydney, Australia. Association for Computational Linguistics.

Ladd, D. R. (1996). *Intonational Phonology*. Cambridge: Cambridge University Press.

Mattys, S. & Jusczyk, P. (2001). Phonotactic cues for segmentation of fluent speech by infants. *Cognition*, *78*, 91–121.

Mattys, S., Jusczyk, P., Luce, P., & Morgan, J. (1999). Phonotactic and prosodic effects on word segmentation in infants. *Cognitive Psychology*, *38*, 465–494.

McDonald, R. & Pereira, F. (2006a). Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of the 10th Conference on Natural Language Learning*, New Brunswick. ACL.

McDonald, R. & Pereira, F. (2006b). Online learning of approximate dependency parsing algorithms. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics,*, pp. 81–88, New Brunswick. ACL.

McDonald, R., Pereira, F., Ribarov, K., & Hajic, J. (2005). Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of the Joint Conference on Human Laguage Technologies and Empirical Methods in Natural Language Processing HLT/EMNLP,*, pp. 523–530, New Brunswick. ACL.

McWhinnie, B. (2005). Item based constructions and the logical problem. In *Proceedings of the Workshop on Psychocomputational Models of Human Language Acquisition. CoNNL-9*, pp. 53–68, New Brunswick. ACL.

Neal, R. & Hinton, G. (1999). A view of the em algorithm that justifies incremental, sparse, and other variants. In M. Jordan (ed.), *Learning in Graphical Models*, pp. 355–368. Cambridge, MA: MIT Press.

Newell, A. (1973). You can't play twenty questions with nature and win. In W. Chase (ed.), *Visual Information Processing*, pp. 283–308. New York NY: Academic Press.

Niyogi, P. (2006). *Computational Nature of Language Learning and Evolution*. Cambridge MA: MIT Press.

Oléron, P. (1953). Conceptual thinking of the deaf. *American Annals of the Deaf*, *98*, 304–310.

Osborne, M. & Briscoe, T. (1997). Learning stochastic categorial grammars. In *Workshop on Computational Natural Language Learning*, pp. 80–87, New Brunswick, NJ. ACL/EACL. Held in conjunction with the 35th Annual Meeting of the Association for Computational Linguistics and the 8th Conference of the European Association for Computational Linguistics, Madrid.

Pereira, F. & Schabes, Y. (1992). Inside-outside reestimation from partially bracketed corpora. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, pp. 128–135. ACL.

Petrick, R. P. A. & Bacchus, F. (2002). A knowledge-based approach to planning with incomplete information and sensing. In M. Ghallab, J. Hertzberg, & P. Traverso (eds.), *Proceedings of the Sixth International Conference on Artificial Intelligence Planning and Scheduling (AIPS-2002)*, pp. 212–221, Menlo Park, CA. AAAI Press.

Petrick, R. P. A. & Bacchus, F. (2004). Extending the knowledge-based approach to planning with incomplete information and sensing. In S. Zilberstein, J. Koehler, & S. Koenig (eds.), *Proceedings of the International Conference on Automated Planning and Scheduling (ICAPS-04)*, pp. 2–11, Menlo Park, CA. AAAI Press.

Quine, W. v. O. (1960). *Word and Object*. Cambridge MA: MIT Press.

Roy, D. (2005). Semiotic schemas: A framework for grounding language in action and perception. *Artificial Intelligence*, *167*, 170–205.

Sakas, W. & Fodor, J. D. (2001). The structural triggers learner. In S. Bertolo (ed.), *Language Acquisition and Learnability*, pp. 172–233. Cambridge: Cambridge University Press.

Siskind, J. (1995). Grounding language in perception. *Artificial Intelligence Review*, *8*, 371–391.

Siskind, J. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, *61*, 39–91.

Steedman, M. (1996). The role of prosody and semantics in the acquisition of syntax. In J. Morgan & K. Demuth (eds.), *Signal to Syntax*, pp. 331–342. Hillsdale, NJ: Erlbaum.

Steedman, M. (2000a). Information structure and the syntax-phonology interface. *Linguistic Inquiry*, *34*, 649–689.

Steedman, M. (2000b). *The Syntactic Process*. Cambridge, MA: MIT Press.

Steedman, M. (2002a). Formalizing affordance. In *Proceedings of the 24th Annual Meeting of the Cognitive Science Society, Fairfax VA, August*, pp. 834–839, Mahwah NJ. Lawrence Erlbaum.

Steedman, M. (2002b). Plans, affordances, and combinatory grammar. *Linguistics and Philosophy*, *25*, 723–753.

Steedman, M. (2007). Compositional semantics of intonation. *Submitted*.

Steedman, M. (2008). Origins of universal grammar in planned action. In M. Christiansen, C. Collins, & S. Edelman (eds.), *Language Universals*, pp. 143–152. Oxford: Oxford University Press. to appear.

Steedman, M. & Baldridge, J. (2006). Combinatory categorial grammar. In K. Brown (ed.), *Encyclopedia of Language and Linguistics*, vol. 2, pp. 610–622. Oxford: Elsevier, 2nd edn.

Steedman, M. & Petrick, R. (2007). Planning dialog actions. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, pp. 265–272. Antwerp, Sept., ACL.

Steels, L. (2004). Constructivist development of grounded construction grammars. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pp. 9–14, Barcelona.

Steels, L. & Baillie, J. C. (2003). Shared grounding of event descriptions by autonomous robots. *Robotics and Autonomous Systems*, *43*, 163–173.

Thompson, C. & Mooney, R. (2003). Acquiring word-meaning mappings for natural language interfaces. *Journal of Artificial Intelligence Research*, *18*, 1–44.

Thornton, R. & Tesan, G. (2006). Categorical acquisition: Parameter setting in universal grammar. *Submitted*.

Tomasello, M. (1999). *The Cultural Origins of Human Cognition*. Cambridge, MA: Harvard University Press.

Villavicencio, A. (2002). The acquisition of a unification-based generalised categorial grammar. Ph.D. thesis, University of Cambridge.

Yang, C. (2002). *Knowledge and Learning in Natural Language*. Oxford: Oxford University Press.

Yang, C. (2006). *The Infinite Gift*. New York NY: Scribner.

Zettlemoyer, L. S., Pasula, H. M., & Kaelbling, L. P. (2005). Learning planning rules in noisy stochastic worlds. In *National Conference on Artificial Intelligence (AAAI)*,. AAAI.

Appendix B

Planning Dialog Actions

Mark Steedman
School of Informatics
University of Edinburgh
Edinburgh EH8 9LW, Scotland, UK
steedman@inf.ed.ac.uk

Ronald P. A. Petrick
School of Informatics
University of Edinburgh
Edinburgh EH8 9LW, Scotland, UK
rpetrick@inf.ed.ac.uk

Abstract

The problem of planning dialog moves can be viewed as an instance of the more general AI problem of planning with incomplete information and sensing. Sensing actions complicate the planning process since such actions engender potentially infinite state spaces. We adapt the Linear Dynamic Event Calculus (LDEC) to the representation of dialog acts using insights from the PKS planner, and show how this formalism can be applied to the problem of planning mixed-initiative collaborative discourse.

1 Introduction

Successful planning in dynamic domains often requires reasoning about sensing acts which, when executed, update the planner's knowledge state without necessarily changing the world state. For instance, reading a piece of paper with a telephone number printed on it may provide the reader with the prerequisite information needed to successfully complete a phone call. Such actions typically have very large, even infinite, sets of possible outcomes in terms of the actual sensed value, and threaten to make search impracticable. There have been several suggestions in the AI literature for how to handle this problem, including Moore (1985); Morgenstern (1988); Etzioni et al. (1992); Stone (1998); and Petrick & Bacchus (2002; 2004).

Stone (2000) points out that the problem of planning effective conversational moves is also a problem of planning with sensing or knowledge-producing actions, a view that is also implicit in early "beliefs, desires and intentions" (BDI) -based approaches (e.g., Litman & Allen (1987); Bratman, Israel & Pollack (1988); Cohen & Levesque (1990); Grosz & Sidner (1990)). Nevertheless, most work on dialog planning has in practice tended to segregate domain planning and discourse planning, treating the former as an AI black box, and capturing the latter in large state-transition machines mediated or controlled via a blackboard or "information state" representing mutual belief, updated by specialized rules more or less directly embodying some form of speech-act theory, dialog game, or theory of textual coherence (e.g., Lambert & Carberry (1991); Traum & Allen (1992); Green & Carberry (1994); Young & Moore (1994); Chu-Carroll & Carberry (1995); Matheson, Poesio & Traum (2000); Beun (2001); Asher & Lascarides (2003); Maudet (2004)). Such accounts often lend themselves to optimization using statistical models (e.g., Singh et al. (2002)).

One of the ostensible reasons for making this separation is that *indirect* speech acts, i.e., achieving coherence via implicatures, abound in conversation. (For instance, Green and Carberry cite studies showing around 13% of answers to Yes/No questions are indirect.) Nevertheless, that very same ubiquity of the phenomenon suggests it is a manifestation of the same planning apparatus as the domain planner, and that it should not be necessary to construct a completely separate specialized planner for dialog acts.

This paper addresses the problem of dialog planning by applying techniques developed in the AI planning literature for handling sensing and incomplete information. To this end, we work with planning domains axiomatized in the language of the

Linear Dynamic Event Calculus (LDEC), but extended with constructs inspired by the knowledge-level conditional planner PKS.

2 Linear Dynamic Event Calculus (LDEC)

The Linear Dynamic Event Calculus (LDEC) (Steedman, 1997; Steedman, 2002) is a logical formalism that combines the insights of the Event Calculus of Kowalski & Sergot (1986), itself a descendant of the Situation Calculus (McCarthy and Hayes, 1969), and the STRIPS planner of Fikes & Nilsson (1971), together with the Dynamic and Linear Logics developed by Girard (1987), Harel (1984), and others.

The particular dynamic logic that we work with here exclusively uses the deterministic "necessity" modality $[\alpha]$. For instance, if a program α computes a function f over the integers, then an expression like "$n \geq 0 \Rightarrow [\alpha](y = f(n))$" indicates that "in any situation in which $n \geq 0$, after every execution of α that terminates, $y = f(n)$." We can think of this modality as defining a logic whose models are Kripke diagrams, where accessibility between situations is represented by events defined in terms of the conditions which must hold before an event can occur (e.g., "$n \geq 0$"), and the consequences of the event that hold as a result (e.g., "$y = f(n)$").

Thus, *actions* (or *events*) in LDEC provide the sole means of change and affect the *fluents* (i.e., properties) of the world being modelled. Like other dynamic logics, LDEC does not use explicit situation terms to denote the state-dependent values of fluents, but instead, chains together finite sequences of actions using a *sequence* operator ";". For instance, $[\alpha_1; \alpha_2; \ldots; \alpha_n]$ denotes a sequence of n actions and $[\alpha_1; \alpha_2; \ldots; \alpha_n]\phi$ means that ϕ must necessarily hold after every execution of this sequence.

One of the novel features of LDEC is that it mixes two types of logical implication. Besides standard (or intuitionistic) implication $\Rightarrow$, LDEC follows Bibel et al. (1989) and others in using *linear* logical implication, denoted by the symbol $\multimap$. Linear implication extends LDEC's representational power and provides a solution to the *frame problem* (McCarthy and Hayes, 1969), as we'll see below.

An LDEC *domain* is formally described by a collection of axioms. For each action α, a domain includes an *action precondition axiom* of the form:

$$L_1 \wedge L_2 \wedge \ldots \wedge L_k \Rightarrow \mathit{affords}(\alpha),$$

where each L_i is a fluent or its negation (we discuss *affords* below), and an *effect axiom* of the form:

$$\{\mathit{affords}(\alpha)\} \wedge \phi \multimap [\alpha]\psi,$$

where ϕ and ψ are conjunctions of fluents or their negations. LDEC domains can also specify a collection of *initial situation axioms* of the form:

$$L_1 \wedge L_2 \wedge \ldots \wedge L_p,$$

where each L_i is a ground fluent literal. Finally, LDEC domains can include a set of background axioms (e.g., for defining the properties of other modal operators), and a set of simple state constraint axioms (e.g., for encoding inter-fluent relationships). We will not discuss the details of these axioms here.

Action precondition axioms specify the applicability conditions of actions using a special *affords* fluent. Effect axioms use linear implication to build certain "update rules" directly into the LDEC representation. In particular, the fluents of ϕ in the antecedent of an effect axiom are treated as consumable resources that are replaced by the fluents of ψ in the consequent when an action α is applied.[1] $\{\mathit{affords}(\alpha)\}$ means that it is not defined whether $\mathit{affords}(\alpha)$ still holds after α. All other fluents are unchanged. Thus, LDEC's use of linear implication builds a STRIPS-style (Fikes and Nilsson, 1971) treatment of action effects into the semantics of the language, which lets us address the frame problem without having to write explicit frame axioms.

Previous work has demonstrated LDEC's versatility as a language for modelling dialog, by introducing notions of speaker/hearer supposition and common ground (Steedman, 2006). This is achieved by defining a new set of modal operators of the form [X], that designate the participants in the dialog and provide a reference point for the shared beliefs that exist between those participants. For instance, [S] and [H] refer to the "speaker" and "hearer", respectively, while $[C_{SH}]$ refers to the common ground between speaker and hearer.[2] Using these modalities

[1] We treat consumed fluents as being made false.

[2] Additional participant modalities can be defined as needed. A set of LDEC background axioms is provided as part of a domain to govern the behaviour of these modalities.

we can write LDEC formulae that capture common propositions that arise in dialog. For instance, [S] p means "the speaker supposes p", [S] [H] p means "the speaker supposes that the hearer supposes p", and [C_{SH}] [X] p means "it is common ground between the speaker and hearer that X supposes p".

In this paper we extend LDEC even further. First, we recognize the need to model *knowledge* in LDEC, which is a necessary prerequisite for planning with sensing actions, including those needed for effective discourse. Second, we require that our extended representation lend itself to tractable reasoning, in order to facilitate a practical implementation. Finally, although LDEC supports classical plan generation through proof (Steedman, 2002), prior work has not addressed the problem of translating LDEC domains into a form that can take advantage of recent planning algorithms for reasoning with incomplete information and sensing. For a solution to these problems we turn to the PKS planner.

3 Planning with Knowledge and Sensing (PKS)

PKS (Planning with Knowledge and Sensing) is a knowledge-level planner that can build conditional plans in the presence of incomplete information and sensing (Petrick and Bacchus, 2002; Petrick and Bacchus, 2004). Unlike traditional approaches that focus on modelling the world state and how actions change that state, PKS works at a much higher level of abstraction: PKS models an agent's knowledge state and how actions affect that knowledge state.

The key idea behind the PKS approach is that the planner's knowledge state is represented using a first-order language. Since reasoning in a general first-order language is impractical, PKS employs a restricted subset of this language and limits the amount of inference it can perform. This approach differs from those approaches that use propositional representations (i.e., without functions and variables) over which complete reasoning is feasible, or works that attempt to represent complete sets of possible worlds (i.e., sets of states compatible with the planner's incomplete knowledge) using BDDs, Graphplan-like structures, clausal representations, or other such techniques.

What makes the PKS approach particularly novel is the level of abstraction at which PKS operates. By reasoning at the knowledge level, PKS can avoid some of the irrelevant distinctions that occur at the world level, which gives rise to efficient inference and plans that are often quite "natural". Although the set of inferences PKS supports is weaker than that of many possible-worlds approaches, PKS can make use of non-propositional features such as functions and variables, allowing it to solve problems that can be difficult for world-level planners.

Like LDEC, PKS is based on a generalization of STRIPS. In STRIPS, the world state is modelled by a single database. In PKS, the planner's knowledge state, rather than the world state, is represented by a set of five databases whose contents have a fixed, formal interpretation in a modal logic of knowledge. To ensure efficient inference, PKS restricts the types of knowledge (especially disjunctions) each database can model. We briefly describe three of these databases (K_f, K_v, and K_w) here.

K_f: This database is like a standard STRIPS database except that both positive and negative facts are stored and the closed world assumption is not applied. K_f can include any ground literal ℓ, where $\ell \in K_f$ means "ℓ is known". K_f can also contain knowledge of function values.

K_v: This database stores information about function values that will become known at execution time, such as the plan-time effects of sensing actions that return numeric values. During planning, PKS can use K_v knowledge of finite-range functions to build multi-way conditional branches into a plan. K_v function terms also act as "run-time variables"—placeholders for function values that will only be available at execution time.

K_w: This database models the plan-time effects of "binary" sensing actions. $\phi \in K_w$ means that at plan time the planner either knows ϕ or knows $\neg\phi$, and that at execution time this disjunction will be resolved. PKS uses such "know-whether" facts to construct binary conditional branches in a plan.

PKS also includes a database (K_x) of known "exclusive-or" disjunctions and a database (*LCW*) for modelling known instances of "local closed world" information (Etzioni et al., 1994).

Actions in PKS are modelled as queries and updates to the databases. *Action preconditions* are specified as a list of *primitive queries* about the state

of the databases: (i) Kp, is p known to be true?, (ii) $K_v t$, is the value of t known?, (iii) $K_w p$, is p known to be true or known to be false (i.e., does the planner know-whether p)?, or (iv) the negation of (i)–(iii). *Action effects* are described by a set of STRIPS-like *database updates* that specify the formulae to be added to and deleted from the databases. These updates capture the changes to the planner's knowledge state that result from executing the action.

Using this representation, PKS constructs plans by applying actions in a simple forward-chaining manner: provided an action's preconditions are satisfied by the planner's knowledge state, an action's effects are applied to form a new knowledge state. Conditional branches can be added to a plan provided the planner has K_w or (particular types of) K_v information. For instance, if the planner has K_w information about a formula p then it can add a binary branch to a plan. Along one branch, p is assumed to be known while along the other branch $\neg p$ is assumed to be known. PKS can also use K_v information to denote certain execution-time quantities in a plan. Planning continues along each branch until all branches satisfy the goal.

4 Planning Speech Acts with LDEC/PKS

Our approach to planning dialog acts aims to introduce certain features of PKS within LDEC, with the goal of generating plans using the PKS framework. In this paper we primarily focus on the representational issues concerning LDEC, and simply sketch our approach for completing the link to PKS.

The most important insight PKS provides is its action representation based on simple *knowledge primitives*: K/K_f "know", K_v "know value", and K_w "know whether". In particular, PKS's tractable treatment of this information—which underlies its databases and queries—is essential to its ability to build plans with incomplete knowledge and sensing.

In order to model similar conditions of incomplete information in LDEC, we introduce a set of PKS-style knowledge primitives into LDEC in the form of *knowledge fluents* (Demolombe and Pozos Parra, 2000). Knowledge fluents are treated as ordinary fluents but are understood to have particular meanings with respect to the knowledge state. For instance, in our earlier example of reading a piece of paper with a telephone number printed on it, we could use a knowledge fluent $KhavePaper$ to indicate that an agent knows it has the required piece of paper, $K_v phoneNumber$ to represent the result of reading the phone number from the paper (i.e., the agent "knows the value of the phone number"), and $K_w connected$ to denote the result of actually dialling the phone number (i.e., the agent "knows whether the call connected successfully").

In a dialog setting, we must also ground all knowledge-level assertions to particular participants in the dialog, or to the common ground. Otherwise, such references will have little meaning in a multi-agent context. Thus, we couple speaker/hearer modalities together with knowledge fluents to write LDEC expressions like $[\mathrm{S}]\,Kp$ — "the speaker knows p", $[\mathrm{H}]\,K_v t$ — "the hearer knows the value of t", or more complex expressions like $[\mathrm{C_{SH}}]\,[\mathrm{H}]\,K_w p$ — "it's common ground between the speaker and hearer that the hearer knows whether p".

Although we treat knowledge fluents as ordinary fluents in LDEC, we retain their knowledge-level meanings with respect to their use in PKS. Thus, knowledge fluents serve a dual purpose in LDEC. First, they act as queries for establishing the truth of particular knowledge-level assertions (e.g., an action precondition axiom like $[\mathrm{X}]\,Kp \Rightarrow \mathit{affords}(\alpha)$ means "if X knows p then this affords action α"). Second, they act as updates that specify how knowledge changes due to action (e.g., an effect axiom like $\{\mathit{affords}(\alpha)\} \multimap [\alpha][X]K_v t$ means "executing α causes X to come to know the value of t"). This correlation between LDEC and PKS is not a coincidence but one, we hope, that will let us use PKS as a target planner for LDEC domains.

We illustrate our LDEC extensions in the following domain axiomatization, which is sufficient to support planning with dialog acts.

4.1 Background Axioms

(1) $[\mathrm{X}]\,p \Rightarrow p$ Supposition Veridicality

(2) $[\mathrm{X}]\,\neg p \Rightarrow \neg\,[\mathrm{X}]\,p$ Supposition Consistency

(3) $\neg\,[\mathrm{X}]\,p \Rightarrow [\mathrm{X}]\,\neg\,[\mathrm{X}]\,p$ Negative Introspection

(4) $[\mathrm{C_{SH}}]\,p \Leftrightarrow ([\mathrm{S}]\,[\mathrm{C_{SH}}]\,p \wedge [\mathrm{H}]\,[\mathrm{C_{SH}}]\,p)$ Common Ground

(5) $[X]\,[C_{XY}]\,p \Rightarrow [X]\,p$
Common Ground Veridicality

4.2 Initial Facts

(6) a. "I suppose Bonnie doesn't know what train I will catch"
b. $[S] \neg [B]\, K_v train$

(7) a. "If I know what time it is, I know what train I will catch."
b. $[S]\, K_v time \Rightarrow [S]\, K_v train$

(8) a. "I don't know what train I will catch."
b. $[S] \neg K_v train$

(9) a. "I suppose you know what time it is."
b. $[S]\,[H]\, K_v time$

(10) a. "I suppose it's not common ground that I don't know what time it is."
b. $[S] \neg [C_{SH}] \neg [S]\, K_v time$

4.3 Rules

(11) a. "If X supposes p, and X supposes p is not common ground, X can tell Y p"
b. $[X]\,p \wedge [X] \neg [C_{XY}]\,p$
$\Rightarrow \mathit{affords}(\mathit{tell}(X, Y, p))$

(12) a. "If X tells Y p, Y stops not knowing it and starts to know it."
b. $\{\mathit{affords}(\mathit{tell}(X, Y, p))\} \wedge \neg [Y]\,p$
$\multimap [\mathit{tell}(X, Y, p)]\,[Y]\,p$

(13) a. "If X doesn't know p and X supposes Y does, X can ask Y about it."
b. $\neg [X]\,p \wedge [X]\,[Y]\,p$
$\Rightarrow \mathit{affords}(\mathit{ask}(X, Y, p))$

(14) a. "If X asks Y about p, it makes it common ground X doesn't know it"
b. $\{\mathit{affords}(\mathit{ask}(X, Y, p))\}$
$\multimap [\mathit{ask}(X, Y, p)]\,[C_{XY}] \neg [X]\,p$

Axioms (1) – (5) capture a set of standard assumptions about speaker/hearer modalities and common ground. In (3), we assume the presence of a negative introspection axiom, however, we do not require its full generality in practice.[3]

Axioms (6) – (10) specify a number of initial facts about speaker/hearer suppositions. In particular, (10) asserts a speaker supposition about common ground that illustrates the types of conclusions we typically require. These facts also include two K_v knowledge fluents, $K_v train$ and $K_v time$. As in PKS, these fluents act as placeholders for the values of known functions that can map to a wide range of possible values, but whose definite values may not be known at plan/reasoning time.

Rules (11) – (14) encode action precondition and effects axioms for two speech acts, *ask* and *tell*.

Using this axiomatization, we consider the task of constructing two dialog-based plans, as a problem of planning through proof.

4.4 Planning a Direct Speech Act

Goal: I need Bonnie to know which train I'll catch.

By speaker supposition, the hearer knows what time it is:

(15) $\Rightarrow [H]\, K_v time$ (*9b*); (*1*)

The speaker doesn't know what time it is:

(16) $\Rightarrow \neg [S]\, K_v time$ (*8b*); (*2*); (*7b*)

By speaker supposition, Bonnie doesn't know what train the speaker will catch:

(17) $\Rightarrow \neg [B]\, K_v train$ (*6b*); (*1*)

The speaker supposes it's not common ground with Bonnie as to what train the speaker will catch:

(18) $\Rightarrow [S] \neg [C_{SB}]\, K_v train$ (*8b*); (*2*); (*5*); (*3*); (*4*)

The situation affords $\mathit{ask}(S, H, K_v time)$:

(19) $\Rightarrow \mathit{affords}(\mathit{ask}(S, H, K_v time))$ (*16*); (*9b*); (*13b*)

After applying $\mathit{ask}(S, H, K_v time)$:

(20) $\Rightarrow [C_{SH}] \neg [S]\, K_v time$ (*19*); (*14b*)

The situation now affords $\mathit{tell}(H, S, K_v time)$:

(21) $\Rightarrow \mathit{affords}(\mathit{tell}(H, S, K_v time))$
(15); (20); (4); (5); (11*b*)

After applying $\mathit{tell}(H, S, K_v time)$:

(22) $\Rightarrow [S]\, K_v time$ (*21*); (*16*); (*12b*)

—which means I know what train I will catch:

(23) $\Rightarrow [S]\, K_v train$ (*22*); (*7b*)

The situation now affords $\mathit{tell}(S, B, K_v train)$

(24) $\Rightarrow \mathit{affords}(\mathit{tell}(S, B, K_v train))$ (*23*); (*18*); (*11b*)

After applying $\mathit{tell}(S, B, K_v train)$:

[3] The weaker property $[X] \neg p \Rightarrow [X] \neg [C_{XY}]\, p$ (which also follows from negative introspection) will typically suffice.

(25) $\Rightarrow$ [B] K_v*train* (*24*); (*17*); (*12b*)

4.5 Planning an Indirect Speech Act

The original situation also affords telling the hearer that I don't know the time:

(26) $\Rightarrow$ [S] $\neg$ [S] K_v*time* (*8b*); (*2*); (*7*); (*3*)

(27) $\Rightarrow$ [S] $\neg$ [C_{SH}] $\neg$ [S] K_v*time* (*10*)

(28) $\Rightarrow$ *affords*(*tell*(S, H, $\neg$ [S] K_v*time*)) (26); (27); (11*b*)

After saying "I don't know what time it is"—that is, applying the action *tell*(S, H, $\neg$ [S] K_v*time*),

(29) $\Rightarrow$ [C_{SH}] $\neg$ [S] K_v*time* (*14b*)

Since (29) is identical to (20), the situation again affords *tell*(H, S, K_v*time*), and the rest of the plan can continue as before.

Asking the time by saying "I don't know what time it is" would usually be regarded as an indirect speech act. Under the present account, both "direct" and "indirect" speech acts have effects that change the same set of facts about the knowledge states of the participants. Both involve inference. In some sense, there is no such thing as a "direct" speech act. In that sense, it is not surprising that indirect speech acts are so widespread: *all* speech acts are indirect in the sense of involving inference. Crucially, the plan does not depend upon the hearer identifying the fact that the speaker's utterance "I don't know what time it is" had the illocutionary force of a request or question such as "What time is it?".

From an axiomatic point of view, the above examples illustrate that the reasoning required to achieve the desired conclusions is straightforward—in most cases only direct applications of the domain axioms are used. Most importantly, we do not need to resolve knowledge-level conclusions like K_v*train* at this level of reasoning and, thus, do not require standard axioms of knowledge to reason about the formulae *within* the scope of $K/K_v/K_w$.

Direct manipulation of fluents like K_v*train* means that we can manage knowledge and sensing actions in a PKS-style manner in our account. For instance, the above plans result in the conclusion [S] K_v*time* as a consequence of the *ask* and *tell* actions. The particular effect of "coming to know the value" of *time* means that we should treat these actions as sensing actions. At the knowledge-level of abstraction, the effects of *ask* and *tell* are no different than the effect produced by reading a piece of paper to come to know a telephone number in our earlier example. This PKS-style use of knowledge fluents also opens up the possibility of constructing conditional plans and, ultimately, planning with PKS itself.

4.6 On So-called Conversational Implicature

The fact that we distinguish speaker suppositions about common ground from the hearer suppositions themselves means that we can include the following rules parallel to (11) and (12) without inconsistency:

(30) a. "X can always say p to Y"
b. $\Rightarrow$ *affords*(*say*(X, Y, p))

(31) a. "If X says p to Y, and Y supposes $\neg p$, then Y continues to suppose $\neg p$, and supposes that $\neg p$ is not common ground."
b. {*affords*(*say*(X, Y, p))} $\wedge$ [Y] $\neg p$
$\multimap$ [*say*(X, Y, p)][Y]$\neg p \wedge$ [Y] $\neg$ [C] $\neg p$

Speakers' calculations about what will follow from making claims about hearers' knowledge states extend to what will follow from making *false* utterances. To take a famous example from Grice, suppose that we both know that you have done me an unfriendly turn:

(32) a. "I know that you are not a good friend"
b. [S] $\neg$*friendship*(h) = *good*

(33) a. "You know that you are not a good friend"
b. [H] $\neg$*friendship*(h) = *good*

After applying *say*(S, H, *friendship*(h) = *good*), say by uttering the following:

(34) You're a fine friend!

the following holds:

(35) $\Rightarrow$ [H] $\neg$*friendship*(h) = *good*
$\wedge$ [H] $\neg$ [C] $\neg$*friendship*(h) = *good* (*32*); (*33*); (*31b*)

One might not think that getting the hearer to infer something they already know is very useful. However, if we assume a mechanism of attention, whereby things that are inferred become salient, then we have drawn their attention to their trespass. Moreover, the information state that we have brought them to is one that would normally suggest,

via rules like (11) and (12), that the hearer should tell the original speaker that they are not a fine friend. Of course, further reflection (via similar rules we pass over here) is likely to make the hearer unwilling to do so, leaving them few conversational gambits other than to slink silently and guiltily away. This of course is what the original speaker really intended.

4.7 A Prediction of the Theory

This theory explains, as Grice did not, why this trope is asymmetrical: the following is predicted to be an ineffectual way to make a hearer pleasantly aware that they have acted as a good friend:

(36) #You're a lousy friend!

It is counterproductive to make the hearer think of the key fact for themselves. Moreover, there is no reason for them not to respond to the contradiction. Unlike (34), this utterance is likely to evoke a vociferous correction to the common ground, rather than smug acquiescence to the contrary, parallel to the sheepish response evoked by (34).

5 Discussion

We have presented a number of toy examples in this paper for purposes of exposition: scaling to realistic domains will raise all the usual problems of knowledge representation that AI is heir to. However, the update effects (and side-effects) of discourse planning that we describe are general-purpose. They are entirely driven by the knowledge state, without recourse to specifically conversational rules, other than some very general rules of consistency maintenance in common ground. There is therefore some hope that conversational planning itself is of low complexity, and that any domain we can actually plan in, we can also plan conversations about.

According to this theory, illocutionary acts such as questioning and requesting are discourse subplans that are emergent from the general rules for maintaining consistency in the common ground and for manipulating knowledge-level information, such as the K_v formulae in our examples. Of course, for practical applications that require efficient execution, we can always memoize the proofs of such frequently-used sub-plans in the way that is standard in Explanation-Based Learning (EBL). For instance, by treating action sequences as "compound" actions in the planning process, we would be in effect compiling them into a model of dialog state-change of the kind that is common in practical dialog management. More importantly, the present work offers a way to derive such models automatically from first principles, rather than laboriously constructing them by hand.

In contrast to approaches that reject the planning model on complexity grounds, e.g., (Beun, 2001), our choice of a planner with limited reasoning capabilities and knowledge resources—conditions often cited as underlying human planning and dialog—aims to address such concerns directly. Furthermore, the specialized rules governing speech act selection in alternate approaches can always be adopted as planning heuristics guiding action choice, if existing planning algorithms fail to produce sufficient plans.

We have also argued that LDEC, extended with PKS-style knowledge primitives, is sufficient for planning dialog actions. Although we have motivated a correspondence between LDEC and PKS, we have not described how PKS planning domains can be formed from LDEC axioms. While some of the mechanisms needed to support a translation already exist—the compilation of LDEC rules into PKS queries and database updates is straightforward and syntactic—we have yet to extend PKS's inference rules to encompass speaker/hearer modalities, and formally prove the soundness of our translation. We are also exploring the use of PKS's *LCW* database to manage common ground as a form of closed world information. (For example, if a participant X cannot establish p as common ground then X should assume p is not common ground.) Finally, we require a comprehensive evaluation of our approach to assess its feasibility and scalability to more complex dialog scenarios. Overall, we are optimistic about our prospects for adapting PKS to the problem of planning dialog acts.

Acknowledgements

The work reported in this paper was partially funded by the European Commission as part of the PACO-PLUS project (FP6-2004-IST-4-27657), and by the NSF under grant number NSF-IIS-0416128.

References

Nicholas Asher and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press, Cambridge.

Robbert-Jan Beun. 2001. On the generation of coherent dialogue. *Pragmatics and Cognition*, 9:37–68.

Wolfgang Bibel, Luis Farinas del Cerro, B. Fronhfer, and A. Herzig. 1989. Plan generation by linear proofs: on semantics. In *German Workshop on Artificial Intelligence - GWAI'89*, volume 216 of *Informatik-Fachberichte*, Berlin. Springer Verlag.

Michael Bratman, David Israel, and Martha Pollack. 1988. Plans and resource-bounded practical reasoning. *Computational Intelligence*, 4:349–355.

Jennifer Chu-Carroll and Sandy Carberry. 1995. Response generation in collaborative negotiation. In *Proceedings of ACL-95*, pages 136–143. ACL.

Philip Cohen and Hector Levesque. 1990. Rational interaction as the basis for communication. In Philip Cohen, Jerry Morgan, and Martha Pollack, editors, *Intentions in Communication*, pages 221–255. MIT Press, Cambridge, MA.

Robert Demolombe and Maria del Pilar Pozos Parra. 2000. A simple and tractable extension of situation calculus to epistemic logic. In *Proceedings of ISMIS-2000*, pages 515–524.

Oren Etzioni, Steve Hanks, Daniel Weld, Denise Draper, Neal Lesh, and Mike Williamson. 1992. An approach to planning with incomplete information. In *Proceedings of KR-92*, pages 115–125.

Oren Etzioni, Keith Golden, and Daniel Weld. 1994. Tractable closed world reasoning with updates. In *Proceedings of KR-94*, pages 178–189. Morgan Kaufmann Publishers.

Richard Fikes and Nils Nilsson. 1971. Strips: a new approach to the application of theorem proving to problem solving. *Artificial Intelligence*, 2:189–208.

Jean-Yves Girard. 1987. Linear logic. *Theoretical Computer Science*, 50:1–102.

Nancy Green and Sandra Carberry. 1994. A hybrid reasoning model for indirect answers. In *Proceedings of ACL-94*, pages 58–65. ACL.

Barbara Grosz and Candace Sidner. 1990. Plans for discourse. In Philip Cohen, Jerry Morgan, and Martha Pollack, editors, *Intentions in Communication*, pages 417–444. MIT Press, Cambridge, MA.

David Harel. 1984. Dynamic logic. In Dov Gabbay and F. Guenthner, editors, *Handbook of Philosophical Logic*, volume II, pages 497–604. Reidel, Dordrecht.

Robert Kowalski and Maurice Sergot. 1986. A logic-based calculus of events. *New Generation Computing*, 4:67–95.

Lynn Lambert and Sandra Carberry. 1991. A tripartite plan-based model of dialogue. In *Proceedings of ACL-91*, pages 47–54. ACL.

Diane Litman and James Allen. 1987. A plan recognition model for subdialogues in conversation. *Cognitive Science*, 11:163–200.

Colin Matheson, Massimo Poesio, and David Traum. 2000. Modeling grounding and discourse obligations using update rules. In *Proceedings of NAACL 2000, Seattle*.

Nicolas Maudet. 2004. Negotiating language games. *Autonomous Agents and Multi-Agent Systems*, 7:229–233.

John McCarthy and Patrick Hayes. 1969. Some philosophical problems from the standpoint of artificial intelligence. In Bernard Meltzer and Donald Michie, editors, *Machine Intelligence*, volume 4, pages 473–502. Edinburgh University Press, Edinburgh.

Robert Moore. 1985. A formal theory of knowledge and action. In Jerry Hobbs and Robert Moore, editors, *Formal Theories of the Commonsense World*, pages 319–358. Ablex, Norwood, NJ. Reprinted as Ch. 3 of (Moore, 1995).

Robert Moore. 1995. *Logic and Representation*, volume 39 of *CSLI Lecture Notes*. CSLI/Cambridge University Press, Stanford CA.

Leora Morgenstern. 1988. *Foundations of a Logic of Knowledge, Action, and Communication*. Ph.D. thesis, NYU, Courant Institute of Mathematical Sciences.

Ronald P. A. Petrick and Fahiem Bacchus. 2002. A knowledge-based approach to planning with incomplete information and sensing. In *Proceedings of AIPS-02*, pages 212–221.

Ronald P. A. Petrick and Fahiem Bacchus. 2004. Extending the knowledge-based approach to planning with incomplete information and sensing. In *Proc. of ICAPS-04*, pages 2–11.

Satinder Singh, Diane Litman, Michael Kearns, and Marilyn Walker. 2002. Optimizing dialogue management with reinforcement learning: Experiments with the NJFun system. *Journal of Artifial Intelligence Research*, 16:105–133.

Mark Steedman. 1997. Temporality. In Johan van Benthem and Alice ter Meulen, editors, *Handbook of Logic and Language*, pages 895–938. North Holland/Elsevier, Amsterdam.

Mark Steedman. 2002. Plans, affordances, and combinatory grammar. *Linguistics and Philosophy*, 25:723–753.

Mark Steedman. 2006. Surface compositional semantics of intonation. *In submission*.

Matthew Stone. 1998. Abductive planning with sensing. In *Proceedings of AAAI-98*, pages 631–636, Menlo Park CA. AAAI.

Matthew Stone. 2000. Towards a computational account of knowledge, action and inference in instructions. *Journal of Language and Computation*, 1:231–246.

David Traum and James Allen. 1992. A speech acts approach to grounding in conversation. In *Proceedings of ICSLP-92*, pages 137–140.

R. Michael Young and Johanna D. Moore. 1994. DPOCL: a principled approach to discourse planning. In *Proceedings of the 7th International Workshop on Natural Language Generation*, pages 13–20.