p123. A computational approach to examining linguistic development dynamically in early bilinguals:

Teresa Satterfield[a)] Rick Riolo[b)] Huzefa Khalil[c)]

[a)]Department of Romance Languages, University of Michigan; tsatter@umich.edu

[b)]Center for the Study of Complex Systems, University of Michigan; rlriolo@umich.edu

[c)] Program in Financial Engineering, University of Michigan; huzefak@umich.edu

Introduction

This project involves the design and implementation of a novel computational model of *bilingual first language acquisition* (BFLA). Specifically, we explore computationally the theory of Phonological Bootstrapping (Christophe et al. 1994, Christophe and Dupoux 1996, Morgan and Demuth 1996, Christophe et al. 1997), working within a bilingual context. To date, several bootstrapping approaches have been proposed assuming monolingual acquisition. Phonological Bootstrapping is a two-part phonological/acoustic-phonetic analysis outlining how infants start to acquire mental representations of word forms in the lexicon and early syntactic representations of their native language(s). According to the phonological bootstrapping model, children initially construct pre-lexical representations on the basis of cues to abstract symbols or linguistic objects (words or categories) that come readily available in the input from perceptual properties associated with that symbol/object. In order to find word forms and to also detect word boundaries for 'bootstrapping' into syntax, various sources of perceptual information are argued to be exploited as cues, including prosody, statistical analyses, and general purpose analogy mechanisms. To the degree that syntactic structures are a projection of lexical properties, researchers suggest that the child's initial representations contain prosodically segmented units that are identifiable for each language and roughly correspond to syntactic units.

In the proposed BFLA scenario the primary linguistic input is bilingual, lexicon(s) of content words and syntactic structures to be represented are in 2 languages. Within the general area of BFLA, 4 interrelated issues shape the research paradigm:

•**Description of Pre-verbal Stage:** nature of the early production and processing mechanisms at 0-12 months of BFLA

•**System-Building:** systems differentiated from the "onset" of development vs. a single system maintained until much later

•**Patterns of Language Development:** rate and sequence of BFLA compared to monolingual FLA

•**Early Code-mixing:** nature of infant mixing of linguistic codes

Overview of the Computational Model

In our computational approach, an adaptable analogical search for structures across primary linguistic inputs is employed. The core architecture is largely based on mechanisms from the Mitchell (1993) and Hofstadter (1995) program of Copycat, a cognitive computer model of 'high-perception.' Copycat has had success as a model of perception and analogy in various complex sequence domains (e.g., music, numbers, and letters), but to date it has not been applied to language learning tasks. We introduce a preliminary computational model called **Babycat**, to begin to test BFLA hypotheses. The domain of the Copycat system is predefined and fixed; hence there is no learning. On the contrary, Babycat "lives" in a dynamic domain. As her domain changes, she learns new concepts. **Babycat** is built up from REPAST (REcursive Porous Agent Simulation Toolkit), a software framework for creating agent based simulations using Java.

Details of Babycat

Slipnet.

•Acts as Babycat's long-term memory, in the form of a network of nodes and links, where "concepts/linguistic knowledge" used in formation of structures in working memory (Workspace) are stored.

•Information conveyed via concepts in Slipnet does not stem from a single node, but from a node and an aura of linked activity in neighboring nodes.

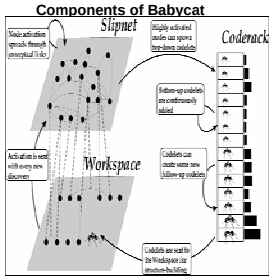

Figure 1. Activity between perceptual loop and conceptual loop

Details of Babycat (cont.)

Slipnet.

•Activity can change based on input, and based on Babycat's current best guess about input, such that the representation can change ('slip') to yield an emergent macro-state interpretation of overall input.

Workspace.

•Data strings are presented here as input into the dynamical area of perception.

•Has a limited capacity (in line with claims that infants initially encode information from a smaller perceptual window than do older learners).

•Input processed in this area ultimately represent perceived relationship between sounds and word forms, sounds and sentence structures; that is, the process of bootstrapping.

Codelets.

•Represent mechanisms of various structural and statistical knowledge that process and modify perceptual interpretations of the input.

•As part of working memory, codelets are placed in Coderack when they are potential processes that represent "currently" competing (and cooperating) hypotheses about how to interpret input. (Coderack is limited in number of codelets).

Temperature.

•Measures the amount of disorganization or entropy in the system's understanding of the given input

•Controls degree of randomness used in making decisions.

Figure 2. Portion of Slipnet is displayed with initial default nodes and links (depicts learner at 0-10 months)

Babycat Program Main Loop

Initialized program: Coderack contains a standard initial population of codelets initially projected to be relevant for every linguistic input presented to the program—if this projection is incorrect, the error will surface only after some codelets have run.

Present input string in the Workspace. Until a sentence has been processed (i.e., representations constructed), do the following:

•Choose a codelet and remove it from the Coderack.

• Run the chosen codelet.

•If N* codelets have run, then (*where N is a parameter currently set at 15):

a. update the Slipnet

b. post bottom-up codelets;

c. post top-down codelets.

Eventually output the constructed representation.

Three Target BFLA Experiments

In seeking responses to our 4 guiding questions on BFLA, we present Babycat's preliminary applications.

•Experiment T1 models a developing linguistic system given consistent and balanced input in Spanish and English, respectively. Objectives: a) to characterize properties of emergent linguistic 'system', and b) to determine what type(s) of pre-lexical representations obtain in this context.

•Experiment T2 introduces only one language into the linguistic input, representing monolingual FLA.

•Experiment T3 implements random input strategies to produce 'unstable' linguistic input as mixed language or code-mixed input.

Experiments are calibrated and evaluated using bilingual language data from the CHILDES database, notably from a longitudinal study (Pérez-Bazán 2002) of 5 Spanish-English bilingual infants, ages 0;8 to 3;0 years, and their caregivers.

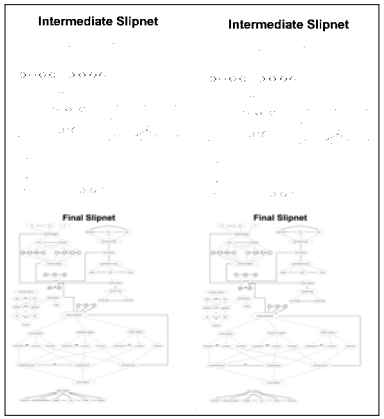

Figure 3. Developing states of a portion of the Slipnet are displayed as screenshots based on proposed Experiments T1:Monolingual English (left column) and T2 Bilingual Spanish-English (right column). The intermediate state depicts learner from 6-12 months, while the 'final' state illustrated here, represents the learner at 24 months.

Preliminary Results and Discussion

It is important to emphasize that Babycat generates "mental" representations of linguistic knowledge in the mind/brain only. It is not a model whose output denotes infant linguistic performance or production.

The Slipnet starts out in the initial state as shown in Figure 2. Figure 3 illustrates the initial activations that have decayed and/or spread in various ways (additional activation has come from codelet actions in the Workspace). The final configuration of the Slipnet indicates what concepts were found to be relevant in the particular input(s). Overall, based on the input sentences ("*Bunny leads a parade"* for monolingual English, followed by "*Da besos Elmo* (Elmo gives kisses)" for bilingual Spanish-English representations), the monolingual representation indicates the acquisition of similar general nodes (linguistic concepts) to that of the bilingual, but with fewer links and activity in the nodes as compared to the Bilingual Slipnet over time.

In seeking responses to our four guiding questions on BFLA, our preliminary results suggest that aspects of development vary substantially according to the domain of linguistic knowledge. We have also made some assumptions concerning the child's ability to use certain early domain-general mechanisms, such as statistical (memory) procedures and prosody features. With these suppositions, the initial Slipnet and Workspace (see Figure 4) are still quite complex and intricate.

While monolingual learners may acquire certain phonemes before their bilingual counterparts, the bilingual's conceptual structures are ultimately 'richer' and allow him/her to manipulate structures in ways that the monolingual cannot achieve. As for the 'Separate or Single System' debate, Babycat weighs in preliminarily showing that, again, different domains of knowledge are continuously developing and changing, and thus it is almost impossible to pinpoint completely separated or completely fused linguistic systems during the initial stages of development. These questions will await more research and refinement of the Babycat model.

In Figure 5, the Slipnet depicting mixed language input is an interesting scenario, as it contains a large amount of ambiguous or identical data across two languages, in this case Spanish and English. The input sentence tested was *"Carla tiene milk,"* with the name *Carla* sharing features in both Spanish and English. The final results are similar to the bilingual Slipnet in that the mixed Slipnet contains numerous active nodes and concepts. The results are similar to the monolingual Slipnet in that the mixed Slipnet contains a smaller number of links and begins from a smaller phoneme base. With further investigation, the representation obtained by Babycat may actually provide insight into the dynamics of creole genesis and the child learner.

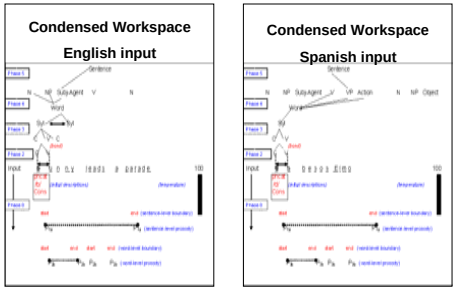

Figure 4. Structures emerging in the Workspace at various levels of development, given English and Spanish input respectively. The temperature, represented by a 'thermometer' on the right, begins at its maximum value of 100 degrees (0 is the minimum), thus initial decision are made fairly randomly, though there are still some biases, even at the highest temperature (e.g., vowels are more salient than consonants, English learners attend to nouns in the input, etc.)

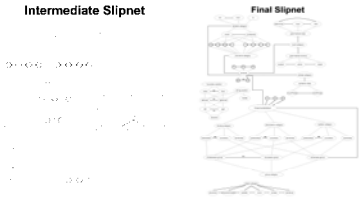

Figure 5. Developing states of a portion of the Slipnet are displayed as screenshots based on proposed Experiment T3 (Mixed language input).

Conclusions

The difficulty in evaluating this model is that it is nearly impossible to prove that a particular linguistic representation is optimal. At this juncture, we see the value of Babycat in carrying out initial experiments that demonstrate proof of concept. By modeling particular conditions and cognitive processes laid out in BFLA analyses and then considering assumptions in the underlying acquisition theory as explicitly as possible, it is possible to suggest the best (human-like) solution, pinpointing theoretical assumptions to determine the degree to which they make each position feasible.

Selected References

Christophe, A., & E. Dupoux. (1996) Bootstrapping lexical acquisition: The role of prosodic structure. ***The Linguistic Review,*** 13, 383-412.

Cristophe, A., E. Dupoux, J. Bertoncini, & J. Mehler. (1994) Do infants perceive word boundaries? An empirical approach to the bootstrapping problem for lexical acquisition. ***Journal Acoustical Society of America*** 95, 1570-80.

Christophe, A., T. Guasti, M. Nespor, E. Dupoux, & B.van Ooyen. (1997) Reflections on phonological bootstrapping: Its role for lexical and syntactic acquisition. ***Language and Cognitive Processes,*** 12, 585-612.

Mitchell, M. (1993) ***Analogy-Making as Perception.*** Cambridge, MA: MIT Press.

Pérez-Bazán, M.J. (2002) Predicting Early Bilingual Development: Towards a probabilistic model of analysis. PhD dissertation, University of Michigan.

Satterfield, T. (1999) ***Bilingual Selection of Syntactic Knowledge: Extending the Principles and Parameters approach.*** Dordrecht: Kluwer.

Sebastián-Gallés, N. & L. Bosch. (2005) Phonology and Bilingualism. In Kroll, J. and A.M.B. de Groot, eds., ***Handbook of Bilingualism: Psycholinguistic Approaches.*** Oxford: Oxford University Press, 63-85.