

MODELING DYNAMICAL LINGUISTIC DEVELOPMENT IN EARLY BILINGUALS: A COMPUTATIONAL APPROACH

1. Project Aims

The proposed project involves the design and implementation of a novel computational model of simultaneous acquisition of two languages from birth, or *bilingual first language acquisition* (BFLA). More 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. 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, and the lexicon(s) of content words and syntactic structures to be built are in two languages.

In first language acquisition research, the concept of "bootstrapping" has been effectively used to explain how children utilize correlations between different aspects of language to infer structure. In our proposed 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, known as Babycat, to begin to test the BFLA hypotheses, and to serve as a springboard for a more advanced version of our current prototype. (Specifications are outlined in Section 4.) To calibrate and also evaluate the experiments that will be implemented in the proposed model, real-world bilingual language data are incorporated into this project from the CHILDES database, most notably data from an original longitudinal study (Pérez-Bazán 2002) of five Spanish-English bilingual infants, ages 0;8 to 3;0 years, and their caregivers.

2. Project Significance

By implementing computational models, this research project can add a new depth to complement previous BFLA efforts: First, the process of formalization—often lacking in studies of bilingual language development—is a crucial endeavor in its own right. Computational models serve as formal versions of possibly ambiguous theories, providing a chance to formulate a clear, consistent, unambiguous version of a "theory." Second, the proposed computational model allows us to carry out novel experiments in which we can predict definite outcomes. Since the model as proposed is able to systematically manipulate variables and closely monitor the consequences of those variations, the ultimate quantitative consequences of a particular hypothesis can be investigated. In this way, we are also able to establish test-retest reliability and predictive validity at a higher level than non-computational approaches. (Experiments detailed in Section 4.) Third,

there is currently an inability for BFLA experimental tools to be shared for collaborative purposes. We view the goal of developing more reusable data-driven language acquisition algorithms as a useful activity and agree that such models should be made available for general research. The proposed model will be developed with an eye toward facilitating the sharing of ideas and resources across several disciplines, such that researchers and students in the social sciences, education, computer science, and linguistics in parallel may utilize our computer program for distinct yet equally relevant purposes.

A coherent base of knowledge is gradually emerging with respect to childhood bilingualism. However, not only is there still a need for more basic research in bilingual development, there is also an increasing call for the inclusion of formal instruments and innovative methodologies using technology to fully explore these research areas (*NICHHD Report on Childhood Bilingualism* 2005). Within the general area of BFLA, four interrelated issues shape the research paradigm:

Description of Pre-verbal Stage. What is the nature of the early production and processing mechanisms at the pre-verbal stage (0-12 months) of BFLA? Given the level of experimentation permissible on very young children, the proposed computational model will be shown to be a useful tool in addressing this question through our innovative examination of the bootstrapping mechanisms.

Unlike the observational and behavioral methodologies of previous BFLA studies (e.g., Bosch and Sebastián-Gallés 1997, 2001, 2003; Burns, Werker and McVie 2003) where the actual object of inquiry (the infant!) is involved in experiments (e.g., “head-turning”, ERP, etc.), the computational model can be executed hundreds of times, systematically modifying the conditions under which it runs and thus exploring the effects of complex variables.

System-Building. Are the systems differentiated “from the onset” (S_1) of development (Genesee 1989, 2000, 2001; De Houwer 1990, 1995; Meisel 1986, 1989, 1990a, 2000), or is a single system maintained until much later (S_n), when the child begins to combine words and employ grammatical morphemes (Swain 1972, Volterra and Taeschner 1978, Vihman 1985)? Our research model will have the capacity to examine this question in detail, by identifying dynamical configurations of the developing mental representations as they emerge at the macro-level (across languages) and the micro-level (within specific linguistic domains such as semantics, phonology, etc.).

With respect to BFLA analyses since the 1970s, most researchers concede that their interpretations of the bilingual data as ‘one- or two-system(s)’ are highly idealized, given that evidence for either position is limited almost exclusively to instances of mixed-language infant production data of single lexical items or morphosyntax in the first word constructions. Until now, these findings have often been justified, as it is argued that no natural methodology exists for examining multiple and interacting linguistic processes in the developing young child. An alternative view is that children begin the language learning task with no system in place (Deuchar and Quay 2000, Tomasello 2000, Vihman 2002). Instead, a system for each language is constructed starting from the increasing number of lexical items accumulated over time. Under this view, the question of ‘one-versus two-systems’ may be insignificant, since the initial state of language acquisition is signaled by the absence of prior (e.g., language-specific) knowledge in domains such as syntax or semantics, and the operative linguistic system is gradually built up.

Patterns of Language Development. Does the rate and sequence of BFLA resemble that of the monolingual? If not, is the difference meaningful? If there are cross-linguistic interactions in the bilingual’s two languages, in which domains do they occur and what mechanisms account for them? The strategy for investigating such questions builds from the previous two issues. Our proposed research can further address this point by demonstrating the ability to model the development of the monolingual

linguistic system in a precise manner that can hold certain variables as constants between the bilingual and monolingual child, in order to extrapolate the relevant information.

There is general consensus that the monolingual and bilingual language acquisition tasks do not differ substantially (Döpke 2000). A serious shortcoming of recent BFLA psycholinguistic studies is their paucity of data covering infants prior to the two-word stage (pre-1;11); findings concerning rates of development are therefore inconclusive. Individual differences between bilinguals are another extremely important dimension of BFLA, however in terms of comparative studies many non-computational experimental approaches have been problematic because of an inability to sufficiently control certain variables. These studies may also overlook interesting aspects that are just too difficult to measure; whereas in our formal model, we will seek to concisely address many of these 'challenging' effects of BFLA such as non-balanced bilingualism (language dominance) and language disorders.

Early Code-mixing (ECM). **Is infant mixing of linguistic codes grammatically constrained? If so, at what point in development do these constraints emerge/become operative?** Our proposed research can formally address the ECM question by encoding code-mixed language as input, and then examining developments within the computational model during the emergence of syntax (through phonological bootstrapping) to determine the types of representations that result.

As yet there is no consensus in the bilingualism literature on what constrains adult code-mixing, and the emerging results on ECM are similarly controversial and inconclusive (Paradis, Nicoladis, and Genesee 2000). If the definition of code-mixing is taken to be "the ability to mix language codes within the same utterance," then this ability presumably distinguishes the bilingual child from her monolingual counterpart, and could be argued to be an integral component in understanding certain representational properties (e.g., syntactic constraints) of bilingualism.

3. Background

A. Foundational Assumptions

We begin with the following assumptions:

-A) A language acquisition model that involves less domain-specific innate structure and more general learning processes is desirable. We uphold the notion that the language architecture (i.e., the cognitive behavior and structural properties of the system's elements) is biologically indicated to the extent that there are innate biases or preferences that trigger processes in each linguistic domain (semantics, syntax, phonology, etc.).

-B) The language learning experience in BFLA differs from that of monolingual acquisition in important ways, even while these groups follow more or less the same processes along more or less the same developmental timeline.

-C) The interfaces of linguistic domains function as grammatical information filters, constraining linguistic data in both one-way and two-way currents.

-D) There is an identifiable and (self)-organizing structure to language representation (e.g., Steels 1997, Satterfield 1999, 2001, among others viewing language as a complex adaptive system, constantly coordinated by numerous internal and external factors), and when particular conditions are met in the developmental process, this dynamical structure constitutes a linguistic "system."

Assumption A is a claim that language acquisition is influenced by a range of cognitive mechanisms, and thus the proposed model does not take one over-arching theoretical perspective exclusively as its foundation. Rather, we seek to provide a principled hybrid perspective between formal and functional explanations. Theoretical investigations of the sort proposed here are very common in many sciences, but still surprisingly controversial in developmental research. This said, the logic is not unknown:

as Meisel's (1990:12) research group states (and Locke 1993 echoes): "...although UG (Universal Grammar) does indeed, according to our hypothesis, function as a 'language acquisition device,' as it used to be called, one cannot hope to explain the patterns of language development unless various mechanisms of language processing and discovery procedures are also taken into account." Assumption B builds on Assumption A by somewhat stating the obvious: 'language acquisition is a complex process.' In developmental research, monolingualism and bilingualism are at best on a scale, moving from complete fluency in two languages to no awareness that other languages exist (Bialystok 2001:8). Assumption C refers to domains of grammar (syntax, semantics, phonology, etc.) as abstract entities that can be relatively non-restrictive when functioning as a self-contained unit. However, given the appropriate interfacing between components, a very rich linguistic system obtains.

Assumption D is a central premise in our program of research. Emergent human linguistic knowledge can be analyzed as a self-organizing system given its complex sets of processes extending across multiple timescales. For this particular inquiry, the complex adaptive system will operate from micro-specifications in the bilingual environment and particular tasks/rules. In turn, the dynamical interactions of these elements can generate macro-structures and collective behaviors in (a) 'global' linguistic system(s).

We therefore assert that part of the task of first language acquisition is to organize knowledge structures such that there is an interaction between the linguistic domains, rather than simply developing the properties within each domain. At state S_0 , the domains function as self-contained units, and at state S_n , they come to interact to varying degrees. A fully operative linguistic system arises once interaction is established between each of the domains, such that there is the ability to generate 'complete' representations of linguistic knowledge (sound, form, and meaning). Prior to the emergence of this capacity, we claim that a general architecture exists, however there are gaps and incomplete knowledge resulting from the developing representations at various stages. As a function of L1 acquisition, learning operations subside when all critical interactions between domains have been installed, thus signaling the maturation (State S_S) of the language system. Preliminary motivation for this proposition comes from cognitive psychology analyses of developmental phases in monolingual children's acquisition of semantics, morphology/syntax, and phonology (Cruttenden 1981; Peters 1986). When a critical mass of given linguistic elements has been acquired, Cruttenden proposes that an innate cognitive mechanism is triggered which attempts to group items and discover relationships among them. Slobin (1985) likewise supports the view that the child accumulates non-unified linguistic knowledge initially, positing that these units come to be systematized via strategies known as Operating Principles.

B. Theoretical Assumptions: Phonological Bootstrapping

The general term "bootstrapping" refers to any process where a simple system activates a more complex system. It is extremely thought-provoking to investigate whether we can arrive at a BFLA solution for early linguistic representation based on the conditions advanced in the following theories. To date, several bootstrapping approaches have been proposed assuming monolingual acquisition:

- *Syntactic (also distributional) bootstrapping*: this theory suggests that grammatical categories can be discovered based on distributional evidence (see e.g., Finch & Chater 1992, Mintz et al. 1995). Mintz et al. (1995), for instance, implement a computational model to show that by monitoring the immediate lexical contexts of words, the similarities of those contexts can be used to cluster lexical items. In an analysis of lexical co-occurrence, Mintz et al. demonstrate that a window of one word to either side of the target lexical item is sufficient to identify nouns and verbs. Specifically, their model reports that there is a 93% probability that an English-speaking child will anticipate a noun following the word "the."

- *Semantic bootstrapping*: under this formulation, word meanings are used as a basis for inferring their grammatical category (Pinker 1987, Bates and MacWhinney 1989). Gentner (1982) argues that nouns have a particularly transparent semantic mapping to the perceptual/conceptual world given that they are object reference terms, and consequently children may use this mapping to designate the category of “noun” in their language and subsequently, acquire syntax.
- *Phonological (prosodic) bootstrapping*: this approach posits that there are phonological or prosodic cues that may point the child to specific linguistic structures (e.g., classes of words, and phrases or clauses) (e.g., Peters 1983, Gleitman et al. 1988, Morgan et al. 1996).

The bootstrapping approaches outlined above emphasize the interaction between domains; that is, the use of information from one independent linguistic domain—the “source” domain—to break into another “target” linguistic domain. Additionally, the correspondence between the source and the target domains is not necessarily a perfect one-to-one mapping; but rather, likely to be only partial. Recently, the notion of “autonomous” bootstrapping has also been introduced which applies within a single, self-contained domain (Cartwright and Brent 1996).

A large body of evidence exists to support the claim that children appear to be sensitive to links between phonological and acoustic-phonetic forms in the acquisition of lexical or syntactic information (see machine learning models by Kelly 1992, 1996; Durieux and Gillis 2001). In the present application of bootstrapping to the bilingual context, we assume that prior to syntactic/distributional or semantic bootstrapping, a two-part model of phonological bootstrapping may allow infants to begin acquiring the lexicon and basic syntax of their first language(s).

According to the phonological bootstrapping model of lexical acquisition, infants build a pre-lexical representation of speech based on perceptual cues in their linguistic environment. This representation is to a large degree language specific, as its purpose is to facilitate the extrapolation of language-specific regularities in the input data. To find word forms, infants a priori do not need to have a lexicon. Instead, they may recruit other sources of information. A growing body of data (e.g., Saffran et al. 1996, Gerken 2002) suggests that infants are able to keep track of various statistical properties of their input by the first year of life: by 6 months, infants are sensitive to typical word shapes in their ambient language(s); by 8 months, infants are capable of noticing distributional regularities (e.g., frequency of certain segments or combinations of segments in certain positions such as word final, frequency of syllable groupings); by 9 months, they are sensitive to the prosody contours (stress patterns in their language, vowel and consonant quality), and phonotactic properties (“legal” versus “illegal” combinations of segments) of native-language words. While the discussion now concerns whether these early discrimination cues can be considered language-learning mechanisms (the same sensitivity to input is found with non-linguistic stimuli and with non-human animals), they nevertheless play a key role in the child's entry into a language system, providing the basis for identifying words, their meanings and grammatical functions, and the kinds of structures they participate in.

In any formulation of how syntactic bootstrapping operates, it appears that a preliminary step entails locating linguistically relevant units in the speech stream; in other words, segmenting the input into ‘chunks’ that correspond to clauses or other linguistic constituents. Thus, the initial prosodically organized pre-lexical representations of the lexicon can also be targeted to break into syntax. Researchers listed above have suggested that the prosodically segmented units roughly correspond to syntactic units. As a general developmental timeline, at 6-9 months, infants are sensitive to prosodic coherence of units of different sizes; at 7 months, infants can parse ongoing speech stream into clause-sized units; whereas at 9 months, infants can parse ongoing speech

into phrase-sized and word-sized units. A resulting representation that the child can extract from phonological information is sentence segmentation and major phrase bracketing as follows:

(1) [ZP [XP the dog] [YP chases the cat]] (Guasti 2004:92)

In this non-hierarchical representation, *ZP* stands for a clause and *XP* and *YP* are major phrases. It is important to note that the phonological bootstrapping hypothesis does not suggest that syntactic structure can be read directly from the phonological representation, or that there is a one-to-one mapping between phonological cues and syntactic units. It only states that an initial segmentation and partial bracketing of the acoustic-phonetic input may get the child started on the process of uncovering syntactic structures by limiting the search space (Morgan 1986).

With respect to the endpoint of bootstrapping, it is speculated that the perceptual, access, or encoding systems are altered or begin having more peripheral effects as the young learner matures. For instance, Newport (1991) proposes that infants initially encode information from a smaller perceptual window than older learners, and the smaller window size yields different, and ultimately, better learning of the more complex form-and-meaning relations.

Theoretical Assumptions: The cognitive architecture (Copycat)

The computational component of our research project implements a modified version of Copycat, a computational model of analogy-making. As a point of departure, we review the main features of Copycat in this section, pinpointing those elements most relevant to our proposed BFLA inquiry.

Copycat is designed to discover “insightful analogies, and to do so in a psychologically realistic way (Hofstadter and Mitchell 1995: 205).” In brief, the Copycat program produces solutions to such problems as:

(2) $abc : abd :: ijk : ?$

(“Initial string *abc* is to Modified string *abd* as Target string *ijk* is to *what?*”) Where the problem solving mechanisms are expected to do the same thing to the sequence ‘*ijk*’ that was done to ‘*abc*’. Hofstadter and Mitchell consider analogy-making to be at the core of “high-level perception” and cognition, which for them are intertwined processes that emerge from the same underlying structures and mechanisms. They identify analogy as basic to recognition and categorization.

In Copycat, high-level perception emerges from the spreading activity of many independent computational agents, called codelets, working on different aspects of an analogy problem simultaneously. They create and destroy temporary perceptual constructs, probabilistically trying out variations to eventually produce a solution. All processing in the model occurs through the collective actions of many codelets over time, rather than on any higher-level centralized process supervising the overall course of events. The model’s high-level “macroscopic” behavior emerges as a consequence of many fine-grained “microscopic” events, thus Copycat is a prime example of “emergent computation.” The codelets rely on an associative network, the Slipnet, built on predefined primitive conceptual components and their associations (long-term memory). (See Figure 1 below.) Changing activation levels of the concepts causes a dynamical conceptual overlap with neighboring concepts. Perceptual structures representative of the Slipnet states are housed in the Workspace, the repository of temporary perceptual structures. The Workspace (also called working memory) is often compared to a blackboard that can be temporarily filled with information.

The formation of an analogy relies on two primary mental processes: representation formation and mapping. Mapping is the process by which objects in each

of the domains are equated with each other because they are perceived as playing a “common role in their contexts,” rather than sharing superficial attributes. In order to perceive that there is a common role between the objects, some kind of mental representation has to be formed that specifies what role each object plays in relation to the others. For example, in the problem stated in (2), a solution such as “*ijl*” would require a mapping between the letter *c* and the letter *k*. In order to make such a correspondence, the objects have to be perceived as playing the same role in both contexts. The representation formed in this case identifies both letters via concept-mapping as the “rightmost” letter in the string, for example.

One of the important aspects to note about the mapping process is that sometimes the objects that are equated do not play the same role in all respects. Conceptual “slippage” is an integral part of the mapping process, and is in line with the partial

correspondence that occurs across domains in the context of phonological bootstrapping. The slippages that occur in these mappings are noted and are used in modifying an initial rule and in generating an answer from the target. Taking another example from the problem domain:

(3) $aabc : aabd :: ijkk : ?$

(“Initial string *aabc* is to Modified string *aabd* as Target string *ijkk* is to *what*?”) A solution such as “*ijll*” demonstrates fluidity and conceptual slippage, where the rightmost letter has slipped to the rightmost group of letters. Copycat is not a deterministic system: as a result, there is no one single “correct” solution; rather, due to its adaptability it can provide a range of solutions to

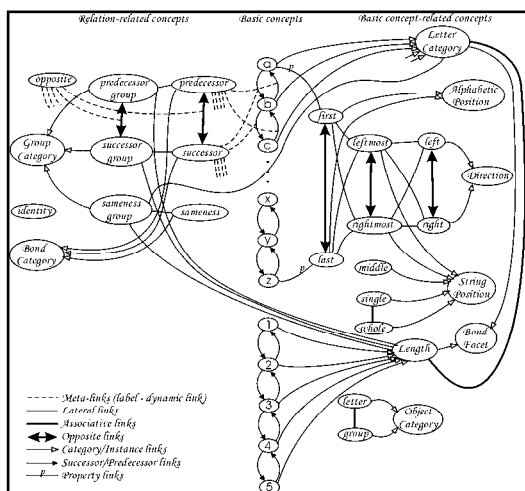


Figure 1. Copycat's Slipnet

the same problem over a number of runs. For certain problems, the solutions judged to be the “best” are even not necessarily the most obvious ones. The fact that Copycat is able to mimic human responses in its domain suggests that it is a plausible and useful model of the mechanisms that drive high-level perception in humans.

To our knowledge, the Copycat program itself has not been applied to any language acquisition problems. Clear models do exist in the literature, however, for how analogy is used in monolingual language learning. The Machine Learning studies in particular contain exemplar-based, instance-based, and other analogical learning algorithms that have been implemented in modeling language acquisition and sentence processing. Algorithms advanced by Skousen (1989) have been developed within linguistics, and are actively used to model irregular morphology among other natural phenomena (also Eddington 2000; see Skousen et al. 2002 for a detailed overview). A drawback of many computational models of analogy-making is that they apply an exhaustive search of the problem domain in order to find the one best possible solution. Such a strategy may be “both computationally infeasible and psychologically implausible in any realistic situation (Hofstadter 1995:286).” Instead of using a brute search, Copycat distinguishes itself by implementing a “parallel terraced scan” in which many (but not all) options are considered in parallel, but to varying degrees. This situation has been achieved by breaking down each task (e.g. building a specific bond), into a series of steps, with each step carrying out a small part of the overall process of mapping the input strings to perceptual and conceptual structures. As Copycat was written to run on serial computers, a form of parallelism is achieved by assigning each step to a different codelet and choosing which codelet to run probabilistically. This approach allows the steps of many different processes to be interwoven, so that the ebb and flow of the activated

processes represents both cooperation and competition between alternative perceptual structures ("hypotheses") to build based on the input.

In the final analysis, the following considerations make Copycat a very promising approach for investigating developmental properties within a BFLA context:

- The notion of analogy-making straightforwardly mirrors the linguistic bootstrapping hypothesis: in both task domains, the goal is to utilize correlations between different aspects to infer structure. In Copycat, this process does not entail "learning" in the sense of acquiring new knowledge or concepts. In analogy-making as well as bootstrapping processes, notions of mapping and representation formation are advanced. Recognition and categorization are also key outcomes in both approaches.
- Copycat follows from the same foundational assumption of complex adaptive systems that defines our proposed research. Recall, we assert that acquisition of first language knowledge is in essence the gradual interaction or linking of domains of grammar. In Copycat, perception emerges from the dynamical spreading activity of many independent codelets, running in parallel. Similarly, bootstrapping approaches emphasize the emergent interaction between independent domains in the use of information from the "source" domain to bootstrap into the "target" linguistic domain.
- In line with our fundamental assumption that language acquisition is influenced by a range of cognitive mechanisms, Copycat is composed from a hybrid architecture, combining both the more traditional paradigm of symbolic Artificial Intelligence and connectionist approaches to cognition (Marshall & Hofstadter 1996).
- Copycat's concept of "slippage" is also compatible with the generally held notion in phonological bootstrapping that the correspondence between the source and the target domains is not necessarily a perfect one-to-one mapping.

4. Prototype Model: Babycat

To recapitulate the major aspects of the proposed research program thus far, we have discussed a working hypothesis for examining early cognitive and linguistic processes at the level of mental representation for BFLA. Although the noted bootstrapping approaches have not represented the bilingual learner, a theory of phonological bootstrapping in early language acquisition could be effective for characterizing pre-lexical BFLA development. Finally, we found the Copycat architecture to be a useful computational instrument for constructing models of cognition and perception. With modifications, we believe that Copycat has the potential to capture the range of behaviors and multiple outcomes commonly encountered in the BFLA experience.

To follow, we describe in the performance of a pilot computational model in phonological bootstrapping, known as Babycat. It was desirable to find a way to translate computationally the fluidity and adaptability of Copycat to high-level perception tasks involving bilingual language acquisition over time. Copycat provides the "architecture" for implementing structures of perception/cognition in a way that allows bottom-up and top-down processes to combine to produce an emergent "solution." Copycat therefore offers a way of viewing both the problem and a resultant answer string.

Babycat builds on and extends the Copycat approach to fluid perception/cognition. In terms of function, Babycat's task is similar to that of Copycat's: to build perceptual structure. Babycat differs from Copycat in that the former is designed to process linguistic inputs and to produce perceptions of a linguistic nature. Babycat abstractly formalizes

those real-world properties of the infant as s/he entertains and ultimately maintains a range of linguistic representations concurrently, as a reflection of the BFLA environment. In sum, we are viewing language acquisition as an abstract puzzle which needs to be solved. The learner's behavior is constrained to the extent that no direct information about the internal workings of the problem or its structure is available *a priori*. However, some initial knowledge must be provided, or the learner would rarely converge to the target language(s). One possible way to attain information is for the learner to be supplied with input and then to observe the results. This point is critical, since it implies that a search procedure is required for solving the language learning "problem." The search technique enables learning to proceed within an overall generate-then-evaluate environment. Thus, a hypothesis is evaluated at a point determined in the search strategy, and it is compared to the target (in this case, the linguistic input). If the target is met, the learning cycle terminates, supplying the evaluated hypothesis as output. If not, the computation (search) continues. Because learning time is limited for infants, the model will impose an upper limit on the number of cycles that can be carried out.

The "solution" produced by Babycat is simply how the model interprets, or represents, the input string in light of Babycat's existing linguistic knowledge structures, as represented by its Slipnet and set of codelets, and the input itself. Since the input is itself the target to be attained, Babycat arrives at the "solution" by trying to build a consistent "perceptual interpretation" (in a broad sense) of the input string. In essence, Babycat tries to assign descriptions to objects, to build objects into higher-level objects (e.g., phonemes into words) and to assign descriptions to those constructed objects, until (if fully successful) it provides a representation for the whole input string as a sentence with the appropriate components, and does so all in the same language. Consequently, another difference in Babycat is that it contains a learning component, such that after it processes a given linguistic input, it will then modify its long term knowledge structures (Slipnet and set of codelets) based on that experience.

It is important to note that Babycat generates "mental" representations (e.g., organization of linguistic knowledge in the mind/brain) only. It is not a model whose output denotes infant linguistic performance or production based on the input data. Of course, the difficulty in evaluating this type of model is that it is nearly impossible to prove that a particular linguistic representation is optimal. At this juncture, we see the value of the Babycat prototype 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 the 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.

Babycat consists of the following components that work in close conjunction:

-*Slipnet*. It is possible to think of the Slipnet as a type of expanded Lexicon per Jackendoff's (1997, 2002) conceptualizations. The Slipnet acts as Babycat's "long-term memory," in the form of a network of nodes and links, where the "meaning" of each concept used in the formation of structures in the Workspace is stored. Babycat contains the relevant atomic concepts (e.g., phonemes, vowels, consonants, word-group, grammatical categories of noun and verb). (Importantly, meaning conveyed via "concepts" in the Slipnet does not stem for a single node, but from a node and an aura of activity in its neighbors. Because this activity can change based on the input, and based on Babycat's current best guess about the input, the concept's meaning can change ("slip") to yield an emergent macro-state "interpretation" of the overall input). This process occurs as Babycat deems that a concept is relevant (when there is an instance of this concept in the given input), it activates its node. If the activation passes a certain threshold, the concept has a probability to be in full activation. Over time, this activation decreases gradually, since concepts fade away if there is no renewed interest in the particular concept. Also, some concepts are more directly perceivable than others, a circumstance reflected by *depth-value*: shallow concepts are perceived immediately, while deep

concepts take longer to get a sufficient activation; however they also lose their activation much more slowly than do shallow concepts.

Correspondingly, links have an ever-changing *conceptual distance*, reflecting the system's current regard for the closeness of the two concepts connected by it. There are a limited number of *link types*, and each type has a concept describing it, for example *identity* or *opposite*. These labels are treated as any other concept, but they have an extra influence on the conceptual distance of their link type. When a concept becomes highly activated, their links shrink, as the current conceptual distance between the two concepts decreases, and vice versa. When *opposite* becomes activated, all opposite concepts in the Slipnet get 'closer' to each other, as diametrical slippages occur more easily while Babycat focuses on 'oppositeness.'

Initially, there is only an activation-free Slipnet. When an input string such as "*Da besos Elmo* (Elmo gives kisses)" is given to the program, the sounds in this string are regarded as *instances* of the phoneme concepts housed in the Slipnet. Because of this relation, the nodes of these sounds in Slipnet become highly activated. These concepts are shallow, since they are easily perceivable, and their activation decays quite fast. However, they also can spread some activation through their links (for example, from 'b' to 'consonant'). The sound instances also get related to their position in the string: the first sound gets linked to 'leftmost', the intermediate sounds to 'middle' and final sounds to 'rightmost'. (These relations are instances, and are not a part of the Slipnet.) Because of these relations, the three concepts also receive high activation due to their relevance, and they can spread it further to 'Left', 'Right' and 'Direction'. These last concepts are deeper, and typically receive little activation at a time (due to their distance), but they also decay more slowly. After a while, activation has spread toward all (at first sight) relevant concepts. Also, when some successive sounds are given, the concept 'successor' will become highly activated, and successor-links will shrink considerably, facilitation flow through these links, making some successive sounds seem relevant, and forcing the program to focus on successorship. The Slipnet thus shapes itself for the language learning problem. After the input has been completely processed, Babycat will then modify its long term knowledge structures (both Slipnet and set of codelets) based on that experience. Learning occurs in this sense that Babycat modifies and acquires certain concepts and new relationships. Thus "learning" in this system includes: modifying the Slipnet by adding nodes, changing default weights on links; or modifying the set of all codelets by adding new codelets or changing functions of codelets.

-*Workspace*. In Babycat, the data strings are presented as input into the dynamical area of perception known as the Workspace. The Workspace has a limited capacity (in line with Newport's (1991) claim that infants initially encode information from a smaller perceptual window than older learners), thus perhaps given the string "*Da besos Elmo* (Elmo gives kisses)", the string "*Da besos*" will be encoded as input to analyzed, based on the infant's perceptual window and the target to be acquired will be "*Da besos Elmo*." While the window may be small, there still can exist an efficient collection of processes used for temporarily storing and manipulating information. The structures that are built in this area represent the perceived relationship between sounds and word forms, sounds and sentence structures, as well as noting which individual elements or groupings are thought to "play the same role" in the utterance. These structures include: *descriptions* (phoneme symbol; distinguishing sound classes such as vowel versus consonant); *object-category* (designates whether input datum is perceived as a grouping or a word); *position-category* (indicates the position where a phoneme is located, such as "leftmost"); *bonds* (mechanisms to establish relationships between adjacent phonemes in each string); *groups* (segmented collections of adjacent phonemes that are linked by common bonds); and *correspondences* (mechanisms for mapping phonemes that are perceived as "playing the same role"). As Babycat generates "perceptual" structures in light of the input string,

these structures function as working hypotheses about how to interpret a portion or all of the input.

d'	a	b	e	s	o	s
Phoneme	Phoneme	Phoneme	Phoneme	Phoneme	Phoneme	Phoneme
L	M	M	M	M	M	R

Figure 2. Portion of ungrouped input string with Description in initial Babycat Workspace

The Workspace is filled with *instances* of the concepts in the Slipnet. Early codelets will link these instances up with their Slipnet counterparts, by attaching *descriptions* to them. When more codelets are recruited, they tend to start working on the most salient instances (A detailed description follows below on codelets.). The *salience value* of an instance is a total of the number of descriptions it has, the activation of the Slipnet nodes it is connected to via these descriptions, and its unhappiness. A concept is *unhappy* when it is hardly used in the existing structures, or is badly integrated. Eventually, the codelets bind some instances into the first conceptual structures. What these structures look like, depends entirely on the codelets, and is as such, very problem-dependent. Descriptions are an exception: they are a fundamental part of the architecture. (We assume in the prototype model that the system has already been trained to ‘perceive’ individual phonemes or phones with a distinctive feature structure, analogous to Gerken’s (2002) phonetic categories).

Once a structure is built, it gets a *strength* value. The strength is influenced by the structure's own properties (e.g. the depth of the concepts used), but it also depends on other structures already built (e.g. how well it fits in), and is evaluated by other codelets. The stronger a structure, the easier it can beat rival structures, if necessary. Some structures can in turn be used as parts of other structures, and can receive their own descriptions. As more structures emerge and the Workspace becomes more complex, automatically a drive toward consistency and the use of deep concepts arises, as ‘inappropriate’ structures do not survive for extended periods. From the correspondences built between the initial input (“*Da besos*”) and the target (“*Da besos Elmo*”), a general rule or representation is built that describes the perceived transformation between these representations. Note that for infants, the pre-lexical representation that results will subsequently be adapted to fit with the target, by looking at what conceptual slippages have been noted to occur within the correspondences between the initial and target strings. A rule is then applied to provide ‘instructions (including any slippages)’ for ultimately arriving at the solution (target string).

-*Set of all Codelets.* The inventory of all possible codelets is stored as long term memory in the Slipnet. Codelets represent all possible mechanisms to build and modify “perceptual interpretations” of the input, and to represent ways in which particular inputs and built structures in the Workspace can influence the activation of the Slipnet. Codelets embody various kinds of structural and statistical knowledge. Babycat’s inventory of codelets includes: *Scouts*, who circulate in the Workspace and examine instances or structures that are in their interest. Their only real action is to propose a certain structure and to call in other codelets to continue the process. *Evaluators* look for the type of object that they were called to assess. They are specialized in estimating the promise of a certain structure. When an evaluator deems the structure interesting enough, it calls in another codelet to do the actual “work.” *Builders* subsequently reify the structure accordingly.

-*Special codelets* will do some other specific tasks related to linguistic analysis (e.g., phonotactic codelets establish relationships between adjacent phonemes in each string;

prosody codelets segment prosodically collections of adjacent phonemes; codelets of distributional regularities can look for particular sounds at the end of words, or pairs of sounds occurring in sequence, and then activate other given nodes in the Slipnet; and in the bilingual context, codelets ultimately can represent statistical regularities found in one language versus another). *Pruners* tear some (weak) structures down when the Workspace is not productive.

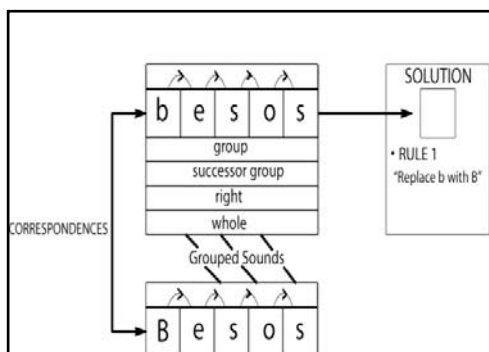


Figure 3: Babycat Workspace: grouping

An example of the codelets in operation: a *description scout* encounters an instance phoneme that could be linked to a Slipnet concept, the scout then calls a *description evaluator* to assess the quality of this possible description. When it appears to be fruitful, a *description builder* and special codelets will be recruited to carry out structure building, as illustrated in Figure 3.

-*Coderack*. As part of working memory, codelets are placed in the Coderack when they are "potential" processes that represent "currently" competing (and cooperating) hypotheses about how to interpret the input. That is, these are processes that have been "nominated" to become active, but may or may not actually become active, based on the collective activity of other codelets, of the slipnet, etc. *Codelets* are chosen probabilistically one at a time to run, based on their urgency values. The urgency values themselves are calculated on a number of factors, including the activation of corresponding nodes in the Slipnet, and how well the structure fits into the existing structures in the Workspace. As the choice of which Codelet to run is a probabilistic decision based on urgency, the Coderack cannot be thought of as a priority procession where the highest priority tasks are run first. Rather, the urgency value dictates the relative speed at which a process is running. For example, codelets seeking instances of a Slipnet concept that is weakly activated in the Slipnet will have a low urgency value and will have to wait a long time before being selected to run. There are three ways in which codelets enter the Coderack. They can be *follow-up codelets*, created by other codelets in order to work further on their findings. Since scouts disappear when they have fulfilled their duty, the Coderack must replenish itself with new scouts, to maintain parallel exploration for new possibilities.

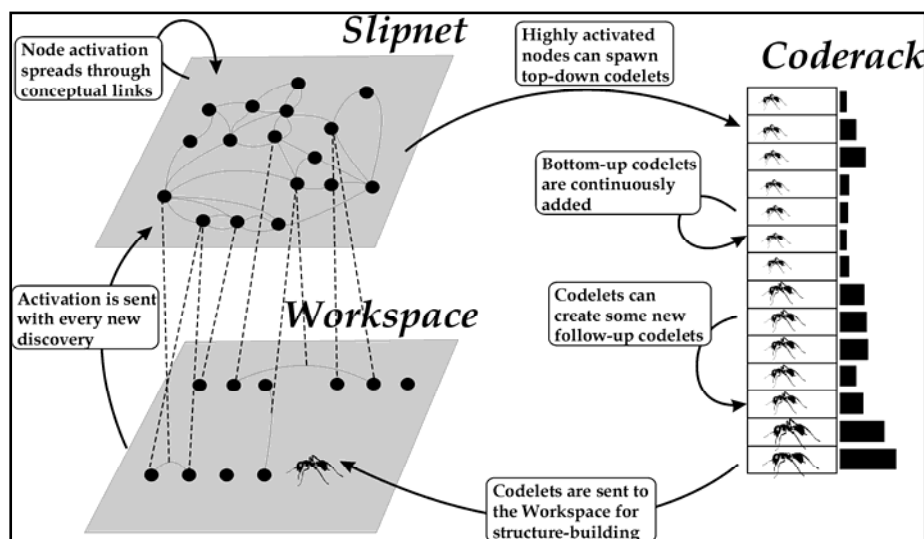


Figure 4: Activity between perceptual loop and conceptual loop

In order to examine development in BFLA, we assume that the learning environment is distributed rather than centralized. We will not impose in advance an organization of grammar(s) (e.g., language-specific knowledge). The goal is to make it possible at various points in the model to determine what general structure of linguistic knowledge is emerging of its own accord, based on the set of possible behaviors/interactions of Babycat. That is, does a fused or differentiated system(s) (or some other structure) emerge early in the Slipnet as the codelets carry out their tasks, and is the result compatible with real-world bilingual child-language development?

In seeking responses to our four guiding questions on BFLA, our research program explores Babycat's preliminary applications. A series of computational experiments will be conducted across several hypotheses. These experiments give us two kinds of knowledge: a) They help us understand how the "theory" works (i.e., how variant implementations of theory and its part work), and the consequences of changing parameters and processes on the behavior of the theoretical system (i.e., the model); b) To the extent the model is a valid representation of the real system, it also suggests how the real system might behave under varying conditions. Thus, the computational experiments that generate interesting changes in behavior suggest possible real-world experiments to carry out, or data to look for from natural experiments), to validate the model or to learn what mechanisms are necessary or not to have a valid model.

All primary input data is taken from the CHILDES data base with bilingual Spanish-English archives, relying heavily on an original longitudinal study (Pérez-Bazán 2002) of five bilingual infants, ages 0;8-3;0 years. Pérez-Bazán's corpus significantly departs from the standard case-study methodology as it sets up a rigorous scientific inquiry utilizing hypotheses and variables to examine in detail a largely equivalent group of infants, comparably acquiring English and Spanish at home. These data are also unique in that statistical analyses are carried out on all utterances recorded. The primary linguistic input directed at the child (parents' speech) which will be incorporated as the linguistic input in Babycat. There is also a statistical analysis of the child's level of bilingualism which can be used to monitor developmental sequencing and to calibrate aspects of the model.

The first experiment (T1) examines the developing linguistic system given consistent and balanced bilingual input of caretaker utterances in Spanish and English, respectively. There are two correlated objectives: a) to characterize the properties of an emergent linguistic 'system', and b) to determine what type(s) of pre-lexical representations obtain in this context. This experiment will provide a preliminary benchmark for exploring the three remaining BFLA questions at length. The second (T2) experiment introduces only one language into the linguistic input, and represents the monolingual learner. The justification for not implementing this scenario as the base case is that while the difference between monolingual and bilingual children needs to be identified, the strategy is often to treat monolinguals as the reference group. Although theoretically interesting, the comparison between monolinguals and bilinguals may not always be the most appropriate comparison, especially when testing developmental or educational measurement instruments. The next three experiments implement random input strategies to produce 'unstable' linguistic input, rather in the form of quantitatively different utterances available in the two respective languages (T3), separate instances of ambiguous input that is difficult to differentiate in the two languages (T4), or consistent mixed languages or code-mixing as input (T5). The sixth experiment goes beyond first language acquisition to explore the inclusion of new second language linguistic knowledge once the system has been in a steady state with one language. This approach may serve to model the language acquisition scenario for many students in the U.S. A final experiment (T7) will manipulate the perceptual cues that are available to infants. Language disorders and learning disabilities are presently difficult to recognize in bilingual populations. Of particular interest is whether BFLA children with language impairments and other challenges can be distinguished from second language learners, for instance, in terms of developmental patterns and rates, and ultimate language proficiency.

Table 1. Experiments and Hypothesized outcomes

Experiment	Hypothesized outcome of Slipnet
T1. L1-L2 Uniform (balanced bilingual input)	large number of linked nodes, many common nodes between L1 and L2 (over time)
T2. L1 Uniform ('balanced' monolingual input)	small number of linked nodes
T3. L1-L2 Non-uniform (dominant bilingual input)	medium number of linked nodes activated, few common nodes between L1 and L2 (over time)
T4. L1'-L2' Uniform (ambiguous forms in bilingual input)	large number of linked nodes activated, most common nodes between L1 and L2 (over time)
T5. L1-L2 Uniform (mixed balanced bilingual input)	large number of linked nodes activated, same nodes between L1 and L2
T6. L1 Monolingual (non-dominant input)	few linked nodes, few common nodes between L1 and L2
T7. L1 Bilingual (impaired input)	outcome dependent on input variable

5. Possible Limitations

While computer models of acquisition and processing may shed light on basic issues in psycholinguistics such as explicit or implicit representation, the amount of innate structure and the amount of linguistic input needed to acquire specific linguistic patterns, they are not always viewed as complementary to observational and behavioral psycholinguistic investigations. Like any research program, the proposed project may be criticized on the grounds of the nature and degree of its underlying descriptions. However, this assessment is not unique to the present project; and we believe that the intrinsic merits of this work overshadow those potential shortcomings. A criticism that can be leveled at computational analyses in general is that every model is a simplification--sometimes a drastic simplification--of the target to be modeled. By all accounts, accuracy (in terms of the number of data points and assumptions built into the model) is important when the aim is prediction, while simplicity is an advantage if the aim is understanding (Axelrod 1997); but the reality is that computational approaches have to satisfy both requirements: a successful predictive model will contribute to understanding at least to some degree, while an explanatory model will always be capable of making some predictions, even if they are not very precise (Gilbert & Troitzsch 1998). Moreover, even if the results obtained from the model match those from the target, there may be some aspects of the target which the model cannot reproduce. A final point concerns the criticism that the outcomes of the model are sensitive to some of the theoretical assumptions grounding it. The notion that computational models are path-dependent in that the results are shaped by the precise initial conditions chosen, does not hold for models representing a complex adaptive system approach. That is, for complex adaptive systems it is not trivial, and sometimes it is impossible to determine what behavior a complex adaptive system will generate, even relatively simple complex adaptive systems like cellular automata (Wolfram 2002). Thus

we must carry out computational experiments to actually determine, *inductively*, the range of behavior that a given complex adaptive systems model (or system) can generate.

6. Implications of the Proposed Research

The four issues of the BFLA research paradigm offer researchers enormous potential for advancing our understanding of child language acquisition and early bilingualism across scientific, social, and applied spheres. There are no precise instruments to account for the actual number of individuals who speak two (or more) languages, yet it is frequently estimated that over half the world's population is bilingual (Grosjean 1982), and that *more* children around the globe are raised as bilinguals than as monolinguals (Tucker 1998). If only in terms of demographics, BFLA studies speak to matters that affect large numbers of individuals worldwide. Despite these statistics, child bilingualism is often treated as a special case or as a departure from the norm. This attitude is especially prevalent in the United States, despite the 2004 U.S. Census estimate that 43% of public school teachers teach bilingual and/or non-English-speaking children. Current BFLA research can be framed in terms of the potentially advantageous effects that bilingualism presents for children's educational success. Recent studies indicate that childhood bilingualism generally promotes an added mental 'dexterity' and creates a deeper reservoir of intellectual abilities. Such findings have immediate and practical relevance for bilingual children, and can be used to influence social and educational policies across the country. In specifically addressing the needs of the bilingual child positive educational outcomes, BFLA research can allow educators to develop reliable tools that identify bilingual learning profiles, pinpointing those areas in learning and literacy where monolingual and bilingual children most differ.

Roman Jakobson, perhaps the most influential linguist of the 20th century, once commented, "Bilingualism is for me the fundamental problem of linguistics." From a theoretical standpoint, a bilingual system inherently reflects 'intricate interactions and configurations' that surpass in complexity those of one single language. From furthering our understanding of BFLA, we broaden our insights into the human capacity for language, and in turn deepen our understanding of the human mind/brain. Furthermore, investigations in BFLA also stand as a rigorous barometer for measuring the viability of our most developed language acquisition theories. There are those who believe that until we have succeeded in establishing a tractable account of the presumably "less complex" theory of the monolingual system, it is premature to compound the linguistic "phenomena" in order to arrive at a theory of the bilingual system. In outlining our research agenda, we will advocate quite the opposite view, given the basic assumption—found in a range of linguistic theories—that early first language learning processes and paths do not differ substantially for monolinguals and bilinguals. Plausibly then, our efforts to understand monolingual linguistic knowledge will more readily succeed through increased attention to intra-theoretic advances in BFLA investigations. As such, it is important to provide descriptively adequate accounts of BFLA acquisition and to more effectively utilize these descriptions to attain explanatory adequacy in the way of developmental theories.