Early Linguistic Interactions: Distributional Properties of Verbs in Syntactic Patterns

Liam Considine

The University of Michigan
Department of Linguistics
April 2012

Advisor: Nick Ellis

Acknowledgements:

I extend my sincerest gratitude to Nick Ellis for agreeing to undertake this project with me. Thank you for cultivating, and partaking in, some of the most enriching experiences of my undergraduate education. The extensive time and energy you invested here has been invaluable to me. Your consistent support and amicable demeanor were truly vital to this learning process.

I want to thank my second reader Ezra Keshet for consenting to evaluate this body of work.

Other thanks go out to Sarah Garvey for helping with precision checking, and Jerry Orlowski for his R code. I am also indebted to Mary Smith and Amanda Graveline for their participation in our weekly meetings. Their presence gave audience to the many intermediate challenges I faced during this project. I also need to thank my roommate Sean and all my other friends for helping me balance this great deal of work with a healthy serving of fun and optimism.

Abstract:

This study explores the statistical distribution of verb type-tokens in verb-argument constructions (VACs). The corpus under investigation is made up of longitudinal child language data from the CHILDES database (MacWhinney 2000). We search a selection of verb patterns identified by the COBUILD pattern grammar project (Francis, Hunston, Manning 1996), these include a number of verb locative constructions (e.g. V in N, V up N, V around N), verb object locative caused-motion constructions (e.g. V Obj on N, V Obj in N), the ditransitive construction (i.e. V Obj Obj) (Goldberg 2006), and the grammatical relations of transitivity and intransitivity (SV, VO) (Ninio 2011). The focus of this investigation is to determine the degree to which language use might optimize construction learning in syntactic development. The search abilities of the CLAN environment (a program specifically designed to analyze transcripts in the CHILDES format) are utilized to extract verb occupancy profiles for the COBUILD argument structures listed above. Each VAC's specifications are translated into CLAN queries that match and return pertinent utterances from the corpus. The data from each VAC search is assembled into a frequency ranked verb type-token distribution. The degree to which each construction exhibits a Zipfian verb type-token distribution is then determined. We calculate mutual information and 1-tau^2 values to gauge the contingency (faithfulness) between each construction and the verbs that occupy it. (Ellis O'Donnell 2011). The nature of children's linguistic input and output is found to be markedly Zipfian, and the verb type-token profile of each construction cannot be accurately predicted from the verb frequency table of the corpus as a whole. This suggests that VAC patterns exhibit a selectional preference for certain items in the verbal repertoire. The final component of this study is an effort to quantify the semantic coherence of the verbs found in each argument-construction. Using WordNet (Miller 2009) and a set of network metrics, the semantic coherence of each VAC's occupancy is assessed. This study shows that the highest frequency VAC occupants are verbs with generic action semantics. These verbs act as strong prototypical exemplars of the constructions underlying functional semantics. This finding is in line with Goldberg's (2006) hypothesis that natural language's recurring Zipfian type-token verb profiles may optimize construction learning by providing a very high frequency exemplar that is prototypical and compatible with the functional semantics of the construction. However, not all constructions have convincingly cohesive lexico-semantic networks and plausible exemplars. SV (intransitive) and VO (transitive) relations are not specialized in this way. Uncovering the organizational properties of a corpus not only tells us about actual language usage, but also provides an important gateway into aspects of speakers' psychologically salient linguistic knowledge. This corpus driven study clarifies the trajectory of linguistic development while also revealing how this follows from important distributional characteristics of children's early linguistic experience.

Table of Contents:

1. Introduction:

The fact that all normal children acquire natural language without any specific training or specially sequenced linguistic data is remarkable. Barring any unusual isolation or cognitive impairments linguistic competence develops universally in all populations of humans. This stands markedly in contrast with the attainment of other comparable cognitive skills. The ability to complete arithmetic calculations or comprehend orthographic texts does not develop in such a robust or predictable manner. These cognitive abilities are learned through deliberate exposure to selected information along with considerable guidance. Despite the substantial variability in a speech community and in each child's set of linguistic experience, they all converge on the same grammatical system.

Modern linguists have an incomplete model of the intricate knowledge that underpins grammar and comprehension, yet every child acquires and implements this information in a matter of years. What about the interface between natural language and children allows the robust acquisition of this complex system? Children develop an adult knowledge state from the linguistic input they receive in conjunction with their genetic endowment. Not everything about the linguistic state of the adult can be found in the input. Like all speech, child directed speech (CDS) is bursts of acoustic disturbances. There are no phonemes, distinctive features, or semantics inherent in these sound waves. The child's perception and processing faculties assign these types of characteristics to the input sounds. Without this native ability the acoustic disturbances would have no linguistic significance. The difference between the input and the fully developed linguistic state is usually attributed to the biological properties of the organisms.

Science uses theoretical abstractions to predict and ultimately explain measurable data. Linguistics is no different. All theories of language hypothesize mechanisms for predicting and explaining grammaticality judgments. Any theory must also account for the infinite creative capacity of natural language. It is the enterprise of the scientific community to examine competing theories and

promote the most capable and economical solutions. Most theories of grammar are about human linguistic knowledge states while other theories focus on speech production and perception. Generative grammar generally assesses unique data using intuition based judgments on grammatical well-formedness. Utterances are carefully selected and compared so that specific linguistic properties can be isolated, experimented upon, described and explained. Construction grammar has arisen from a slightly different way of inspecting linguistic data: primarily the methodologies of corpus linguistics and cognitive grammar.

Empirical corpus studies on language acquisition have been done from both theoretical vantage points. Anat Ninio in her 2011 work *Syntactic Development, its input and output* (*SDIO*) tries to frame her corpus linguistic study between the theoretical devices of the Minimalist Program (Chomsky 1995) and that of Construction grammar (Goldberg 1995). The prominent features of Ninio's work will be contrasted with the studies in Adele Goldberg's 2006 book *Constructions at work*. Reviewing these author's methodology and findings will be important in placing the original content of this study into perspective.

1.1 Theoretical Positions:

Ninio's study is focused on the development of the simplest clausal structures. At the earliest stage of syntactic development Ninio believes children learn two word combinations that make up the core relations of the English clause (subject-verb, verb-object, and verb-indirect object). This study views parental speech both as the input the child receives, and as the immediate goal of the acquisition process. Initially, Ninio leaves the theoretical description of these basic word-combinations open. Nevertheless, she presents the two contemporary theories leading the field. Are simple sentences combinations of atomic categories generated by the Merge operation as assumed in the Minimalist Program (Chomsky 1995) or are they the kind of constructions proposed by Construction Grammars

(Goldberg 1995)? The theoretical concern between Merge couplets or Argument Structure constructions, does not affect the methodology of Nino's project. Ninio's concludes that the highest frequency verbal occupants of the clausal core do not correspond to the prototypical semantic function of the relationships constituents. However, this finding needs to be viewed in context. The research strategy *SDIO* implements is derived from corpus-based linguistics, in that Ninio searches for both the fine details and global characteristics of these syntactic patterns in child and parental speech. The study is based on large pooled speech corpora from the CHILDES database. To achieve the best possible results, Ninio invested a great deal of time in hand parsing and coding the corpora.

SDIO proposes a unique model for the development of linguistic knowledge and syntax. Ninio presents a developmental model where children learn the unique linguistic coding of their speech environment by matching unknown linguistic forms with speakers communicative intent. The interactive context of this linguistic exchange, along with the child's ability to attribute mental states (desires, beliefs, intentions, and knowledge) to other humans, facilitates this process. Children are motivated in choosing which items to learn by the same pragmatic principles that govern adults in choosing which items to say (Ninio 2011: 2). These deeper principles underlying linguistic behavior are left undefined. Following this stance, Ninio predicts that children's English will be, altogether, exceedingly similar to parental speech in its global features. The degree of similarity between child and adult corpora can be used to evaluate the validity of this model of acquisition.

In Ninio's conception, language is a complex network consisting of linguistic items such as words, and speakers who produce words and sentences. Children form new nodes in the existing speech network and make connections to other human nodes by interacting with them and experiencing speech. As children learn to produce syntactic combinations, they are not reinventing or internalizing English. Rather, children link into a network of other speakers who are producing similar combinations (Ninio

2011: 3). Children are adjoined into the network and begin producing the structures created by, and relevant to, their proximate nodes. In this way, no one person can be said to have a complete knowledge the English language. The English language is a composite of a particular section of the human speech network.

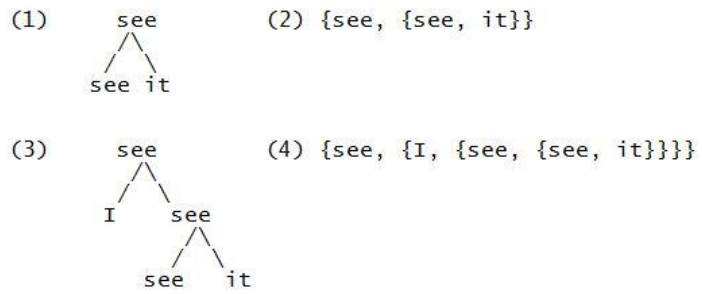
Like other complex networks, language networks are governed by the principle of preferential attachment; children choose the most frequent forms of expression in the input for each communicative need they learn to express. This property causes language networks to retain and enhance the standing global features both before and after the new child user joins. Starting from this network theoretic model, Ninio's research goes against the usual reductionist mode of science. The system is the focus of the study rather than its collective parts. Ninio embraces the reductionist examination of parts if it can help in understanding system-level behavior.

The core syntactic relationships of the English clause: subject-verb (SV), verb-object (VO), and verb-indirect object (VI) are fundamental structures in most theories of syntactic structure. In generative grammar these syntactic relations are meaningless formal patterns in a system autonomous of semantics created by the Merge operation (Ninio 2011: 9). The formal patterns are available for use by certain verbs according to a priori specifications (i.e. lexical features). *Syntactic development: its input and output* (SDIO) uses the corpora of parental and child speech to explore the global features of these syntactic relations.

1.2 Merge

The theoretical apparatus used to generate syntax in the Minimalist Program is Merge (Chomsky 1995). Merge is an asymmetrical process that creates a single syntactic unit from the combination of two previous ones. Merge takes two elements, combines them and assigns a label to the structure thus formed. The label is identical to one of the two elements that are merged.

The Head unit always becomes the label, and also provides the grammatical features to the newly formed element. The Dependent is either the modifier or the complement of the Head. For example, if the Head of a combination is a noun (paint) and the Dependent is a modifying adjective (pink), the combination will be a noun-phrase labeled {paint, {pink, paint}} the resulting structure can occur in the same environments as its Head-noun, such as being the object of a verb. The dependent (pink) specifies some aspect of the Head (paint). Merge is not an arbitrary or random operation; it is always the formal expression of a relation between a Head and Dependent word. There is no definite theory of projection which predicts which element of Merge is the Head or the Dependent. However let's consider the sentence "I see it".



The diagrams labeled (1) and (3) are a tree structure notation of the Merge operation, while (2) and (4) are the bracket notation utilized previously. In (1) *see* is merged with *it.* The Head is the verb *see,* and therefore it is the label of the newly formed constituent. We know the single element formed by this merger of the atomic elements the verb *see* and the pronoun *it* is not some type of noun group, but instead some kind of verbal syntactic molecule. Since Merge is recursive and can apply to its own output, the two member set {see, {see, it}} is merged with the pronoun *I* to create the set {I, {see, {see, it}}}. Again we find the Head of this merger to be the verb *see* for precisely the same reason as before. *See* is projected as the label of the merger between *I* and the two member set {see, {see, it}} because the

it is clear that the top level constituent of this syntactic tree is not some kind of noun but rather some kind of verbal unit. Determining projection in this way is similar to the types of stipulations used in rule based grammars like VP → V NP. There are some semi-useful generalizations that can be made about the projection of labels: principally, if the a lexical item is merging with a label, the lexical item more often than not is the Head projected as the label of the merger. As you can see the example explicated here does not follow this axiom.

The Merge operation has some advantages when it comes to hypothesizing in a developmental domain. In the example above there is a built in reason why the word *it* acts as the Dependent of the Head *see*. Children must accumulate thematic knowledge about lexical items in order for them to have a fully specified lexical entry. If a child does not know that "sit" takes a "sitter" argument (i.e. a person, or some number of people must do the sitting) than they have an incomplete mental representation of the word "sit". Learning lexical semantics is an integral part of knowing a word's syntactic behavior. In this perspective, syntax, and specifically Merge, offer a way to express the 'thing seen' or the 'person doing the sitting' by unifying a separate lexical item with a verb.

"In our system, children start with meaningful words (like *sit*) whose meaning include a requirement for a semantic complement (the identity of the 'sitter'), and syntax is there merely to lead them to the relevant expression." (Ninio 2011: 15).

Merge is not seen as a concrete extension of a word meaning. This view of syntactic development does not require children to extract statistical patterns from multiword utterances in the hope that they will find a useful grammatical combination. Instead, words with syntactic potential (particularly verbs) project some logical-semantic arguments.

At an intuitive level, one can understand verbs as words that describe events in the world. It then follows that verb's arguments are the named individuals that share some principled relationship to that event. Rather than varying arbitrarily or idiosyncratically in the ways they describe events, it is known

that verbs with similar semantic characteristics often exhibit common syntactic characteristics. For example, *break*-type verbs (e.g. *break, crack, rip, shatter*) appear in the middle construction but not in the conative construction or the body-part possessor ascension construction, while *hit*-type verbs (e.g. *bash, hit, kick, pound*) exhibit the opposite pattern. (Levin 1993: 6)

1. [* is used to indicate ungrammaticality]

a. Joshua {hit/*broke} at the vase 'conative'
cf. Carla {hit/broke} the vase

b. Joshua {hit/*broke} Carla on the back '*body-part possessor ascension'*
cf. Joshua {hit/broke} Bill's back

c. The vase {*hit/broke} easily '*middle'*
cf. Joshua {hit/broke} the vase easily

In generative grammar verbs subcategorize syntactically for the required and optional thematic complements. This type of information often called the thematic grid is stored at the lexical level alongside the sound, meaning, and morphological class of the word. In this way Minimalism is radically lexicalist. Syntactic connections are thought to be projected from and grounded in the lexicon. The simplicity of this generative theory is attractive. Much of what acquisition becomes is the learning of individual lexical items and the Merge operation used to combine them. In this analysis there is no need for children to learn abstract rules, phrasal rules, constructions or any other combinatory operation. There is however, some question about how the child knows which constituent is Head and which constituent is Dependent.

Generative grammar requires a theory of how predicates and arguments listed with a verb in the lexicon determine syntactic behavior (i.e. map onto or link with syntactic positions). Theta assignment in generative grammar is an essential element of the theoretical apparatus. The mechanism for theta-assignment involves a number of conceptions. Here is a very brief overview and a short example.

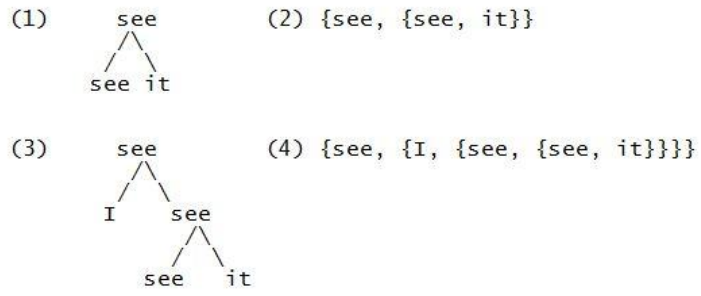
Theta criterion:
Each argument is assigned one and only one theta role.
Each theta role is assigned to one and only one argument.

C-command: x c-commands y iff the first branching node that dominates x dominates y

Government: X governs Y iff X and Y c-command each other.

The Principle of theta-role Assignment:
X assigns a theta-role to Y iff X governs Y.

Paradigmatic examples of theta-assignment:
Object position of a transitive verb in active voice
Subject position of a VP that is headed by a V with an requiring an 'agent' theta-role



Consider again the derivational representation of the sentence "I see it." In this sentence *see* has two theta roles to assign: agent (experience) and theme (stimulus). In (1) as *see* and *it* merge, they come into a relationship of mutual c-command or government. This is the necessary structural relationship for theta-assignment to occur and therefore *see* assigns its 'theme' (i.e. 'object being seen') role to the noun *it*. When merge is applied recursively with the label *see* and pronoun *I* and the two syntactic objects are joined, the label *see* and the pronoun *I* once more enter a relationship of mutual c-command. In this way the external theta-role 'agent' (i.e. 'seer') can be assigned from the label *see* to the pronoun head *I*. The label *see* created by the first merger is not technically the verb *see*. Note that in this particular example,

the agent theta-role is being assigned compositionally from the label *see*, what would have been called V' under X-bar theory.

The notion of Merge is intimately related to the idea that syntactic form is autonomous of semantics. This statement is derived from the fact that core grammatical relations formed by Merge have no association with any particular meaning. We can see the vast array of semantic functions that subjects and direct objects in English show. First let us look at the multiple semantic roles of the grammatical subject.

a. Patient of state: "She is tall"
b. Patient of change: "She is falling asleep"
c. Agent: "She is writing a letter"

Second a similar diversity of semantic roles can be affiliated with the grammatical object

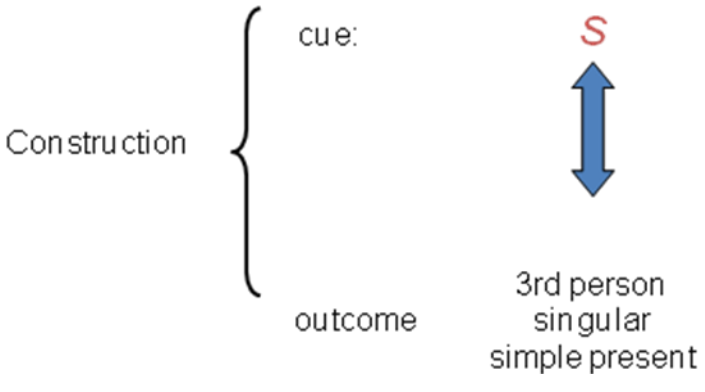
a. Patient of state: "He saw her"
b. Patient of change: "He pushed her"
c. Ablative: "He approached her"
d. Ingressive: "He entered the house"
e. Benefactive: "He built her a house"
f. Dative: "He gave her a book" (Givon 1997)

Here we can see that while the syntactic geometry between SV and VO are consistent in each example, the semantic roles the subject and the direct object play are dynamic.

1.3 Construction Grammar:

There is another leading theory that conceptualizes grammatical relations differently. This theory is called Construction Grammar (Goldberg 1995, Goldberg 2006). From a constructionist position,

language is made up of a structured inventory of constructions which are conventionalized pairings of form and meaning. Usage leads to these constructions becoming entrenched as grammatical knowledge in the speaker's mind, and the degree of entrenchment is directly proportional to the frequency of use (Bybee, 2005; Ellis, 2002a; Langacker, 2000). From a construction grammar stance, each syntactic relationship is a meaningful construction that possesses its own prototypical semantic utility. In Construction Grammar (CXG), many levels of linguistic representation are understood to be constructions. Morphemes (*-s, -ing*), words (*support*, *report*, *the*), idioms (*break a sweat*), and more complex syntactic patterns like the passive (Subj aux VP (PP_{by})) and ditransitive (Subj V Obj_1 Obj_2) are all constructions. In Construction Grammar, language development follows the same cognitive principles as the learning of other categories, schema and prototypes (Cohen & Lefebvre, 2005). Creative linguistic competence emerges from the amalgamation of the memories of all the utterances in a learner's entire history of language use. This consolidation of memories is the fundamental basis of an unconscious knowledge of frequency-biased regularities in linguistic data. (Ellis 2002). Shown below is a diagram representing a simple construction at the morphological level. The cue (or form) of the construction is the suffix *s*, while the outcome (or conventionalized function) of this constructional pair is that of the 3rd person singular simple present.



(Ellis 2009)

One can easily see how in this perspective, a word is defined as a construction. The phonetic or visual form of a word is the symbolic cue, and the related outcome is that item's definition. While this simple example demonstrates the relationship between linguistic form and discourse function, it is still somewhat unclear how constructions are combined to create novel utterances. Consider the sentence "*What did Michael send Miriam?*" under the constructionist account this structure is derived via the application of a number of different constructions. First, each of the five words is a construction; then there is the VP Construction, the NP Construction, a Subject-Auxiliary Inversion Construction, the WH Construction, and the Ditransitive Construction. Without detailing the pure formalities of these rules consider generally what each operation does. The VP and NP Constructions act similar to VP and NP phrase structure rules in that they create larger phrases out of individual words. The WH construction licenses the theme argument of (i.e. the '*what'* being sent) in sentence initial position. The Ditransitive Construction encodes the remaining argument roles to the other constituents based on which slot in the structural pattern they occupy. The theta-role of the constituent immediately after the verb *Michael* is that of the agent and the theta-role of the argument immediately after the verb *Miriam* is that of recipient. If this sentence was not a WH-question, then the WH construction would not have applied and the sentence final argument (i.e. the direct object) would be marked as the theme. The ditransitive construction joins with the action semantics and other subtle notions (manner etc.) specified by the verb. Once the verb *send* fuses with the ditransitive constructions functional semantics (i.e. X CAUSES Y TO RECEIVE Z) the notion of how each person or object stands in relation to one and other (the theta-grid) is known. (Boas, 2010)

Constructions are understood to be learned via induction on the basis of the input along with the general cognitive mechanisms. This sums up what is called the usage-based model of language-acquisition. The primary impetus for construction grammar is to unify linguistic form, learner cognition,

and usage. Importantly constructions cannot be defined purely on the basis of structure, or semantics, or frequency of usage – all of these factors are necessary to formalize and operationalize a construction grammar (Ellis 2011).

As explicated above, basic sentence patterns of language can be understood to involve constructions. Verbs are seen as combining with a syntactic construction (e.g. transitive, intransitive, passive, etc.) from which the argument relation of the structures constituents becomes known. The alternative approach is to assume that the main verb specifies the syntactic location, and number of arguments. In this view, the interpretation of a linguistic pattern is heavily dependent on the verb and the derivational structure of the utterance. The patterns shown in (1) and (2) below do appear to be determined by the lexical specifications of *give* and *put*:

(1) Henry gave Fran the spoon.

(2) George put the record on the turntable.

Give is lexically specified as a three argument verb and is assumed to appear with three complements corresponding to: agent (Henry), recipient (Fran), and theme (the spoon). *Put* is also a three argument verb requiring: an agent (George), a theme (the record), and a location (on the turntable). Both (1) and (2) represent normal cases. However the interpretation and structure of sentence patterns in language cannot always be determined by the independent lexical specifications of the main verb. For example, it is quite implausible to state that the verb *sneeze* has a three-argument sense, yet it appears in pattern (3) below.

(3) "He sneezed the napkin off the table."

Verbs often appear in a wide variety of complement configurations. The examples do not need to be particularly unusual like (3) in order to demonstrate this. In (4) – (8) the verb *slice* occupies a wide array of patterns.

(4) He sliced the bread. (transitive)

(5) Pat sliced the carrots into the salad. (caused motion)

(6) Pat sliced Chris a piece of pie. (ditransitive)

(7) Emeril sliced and diced his way to stardom. (way construction)

(8) Pat sliced the box open. (resultative)

(Goldberg 2006: 7)

In each of these examples *slice* means to cut with a sharp instrument. As Goldberg explains, it is the argument structure constructions (i.e. the constructions used to express the basic clauses) that facilitate the link between surface form and aspects of thematic interpretation. In (4) the transitive construction licenses 'he' as the something acting, and the 'bread' as the thing being acted upon. The interpretation in the other three are, (5) something causing something else to move, (6) someone intending to cause someone to receive something, (7) someone moving somewhere despite obstacles, (8) someone causing something to change state) (Goldberg 1995). This data illustrates the wide variety of complement configuration a single verb can occupy. Generative grammar may require multiple instances of *slice* to be stored in the lexicon each varying in the number of arguments and the specific interpretational role they assume. In construction grammar, there is no need for multiple lexical entries of each verb, because the appropriate interpretational roles are stored in the 'construction'.

Both Goldberg and Ninio bring empirical corpus-based evidence into the evaluation of these competing theoretical descriptions. For the moment, let us examine *SDIO*'s and *Constructions at Work* methodological processes, empirical studies and developmental conclusions.

2.1 Empirical Studies on Construction Learning

Statistical cues are a powerful channel by which initial language learning can begin. Children are able to identify word forms from continuous speech streams based on transitional probabilities between syllables (Saffran, Aslin, and Newport 1996). Statistical cues also enable children to identify syntactic

regularities about categories of words. The presence of an article (*the* or *a*) predicts a noun somewhere later in the speech stream; this facilitates the learning of syntactic phrase boundaries. Goldberg claims that this type of linguistic information is immensely important to language acquisition because "there are no stable formal cues cross linguistically to identify word forms, grammatical categories, or syntactic relations" (Goldberg 2006: 71).

Within a given language, linguists have identified certain formal patterns that correlate strongly with the interpretation of the utterances they appear in. As discussed above, such relations between form and meaning are described many ways. Some conceive of them as linking rules projected from the main verb's lexical specifications (Levin and Rappapport Hovav 1995). While others (like Goldberg) identify them as phrasal form and meaning correspondences that exist independently of particular verbs. The idea, regardless of theoretical preference or terminology, is that certain patterns in language are learnable on the basis of general categorization strategies. The ultimate view is that strong correlations exist between formal linguistic patterns and communicative function. Goldberg presents a broad range of empirical evidence in support of these claims. We must understand how Goldberg relates the cognitive literature on categorical learning and the acquisition of constructions. One part of the larger hypothesis is that constructions which are instantiated (to a high degree) by a single verb are initially easier to learn than those constructions which are typified by many different verbs.

First Goldberg starts with an investigation of speech transcript data from the CHILDES database. (MacWhinney 2000). The corpus used is comprised of data from the Bates, Bretherton, and Snyder developmental studies. In analyzing the mothers' speech, Goldberg found that the use of a particular construction is typically dominated by instances containing one distinct verb. In mothers' speech, *go* was present in 39% (136/353) of the intransitive motion construction's occurrences (Subj V $Obl_{path/loc}$ = X moves $Y_{path/loc}$). In the Bates corpus *Put* made up 38% (99/256) of the instances of the caused-motion

construction (Subj V Obj $Obl_{path/loc}$ = X causes Y to move $Z_{path/loc}$). *Give* was found to occur in 20% (11/54) of the ditransitive constructions (Subj V Obj Obj_2 = X causes Y to receive Z) (Goldberg 2011:76). The same trend noted here in mothers' speech is mirrored in children's early speech. The question arises as to why *put, give,* and *go* are so frequent in the input. The first reason seems to be that *go* and *put* are more frequent than *saunter* or *stack* because they apply to a wider range of arguments, and are therefore relevant in a wider context. The second reason is that many basic verbs designate a fundamental pattern of experience. *Put, give, go,* and *make* are highly neutral in their action semantics, and serve as templates from which to derive the argument structure of an utterance. The frequency of these archetypal verbs is high because they resemble and reinforce the basic function encoded by the constructions they inhabit.

Casenhiser and Goldberg (2005) ran an experiment to test learners' ability to correlate a novel constructional meaning with a novel linguistic form. This is precisely the task that a child faces when naturalistically learning language. In this study a novel pattern involving known nouns was arranged in a non-English word order along with a nonsense verb. The test was a forced choice comprehension task, where the subjects saw two film clips side by side and heard a sentence describing one of the clips. They were then asked to decide which scene the utterance described.

There were a total of five novel verbs and sixteen examples during the pre-test training session. The participants were 51 children with a mean age of 6;4 who were randomly assigned to three conditions; the control, the balanced frequency and the skewed frequency. In the pre-test training only one piece of film was shown with the test phrase. The balanced frequency subjects heard the five novel verbs each with a relatively low token frequency, three novel verbs each occurring four times, and two novel verbs each occurring twice. In the skewed frequency training condition the subjects again heard the five novel verbs, but this time one had an especially high token frequency of eight, while the other

four occurred only twice. The control condition watched the same pre-test training film with the sound turned off. The meaning of the phrasal pattern in training was that of *appearance*. The entity named by the first noun phrase (np1) came to exist in the place named by the second noun phrase (np2). For example the utterance "the sailor$_{np1}$ the pond$_{np2}$ neebod$_{verb}$" was paired with a scene where the sailor sailed onto the pond from out of sight (Goldberg 2006: 80).

As anticipated, the control group did no better than chance in choosing the correct scene during testing. The balanced condition had a statistically significant improvement over the control group in choosing which video matched the content of the linguistic input. This means that they had learned something about the construction's conventionalized form and the way it encoded semantics during the training session. The skewed frequency condition showed a statistically significant improvement in performance over the balanced condition. Goldberg concludes that with less than three minutes of training, both children and adults (the test was also run on undergraduate students) were able to learn a construction form-meaning mapping, and then extend the function of the construction to a novel verb and new scenes. Importantly, the results demonstrate that the high token frequency of a single exemplar does increase the learnability of constructional meaning (Goldberg 2006: 82). Without explaining the methodology in full detail, it is worthy to note that Goldberg and Casenhiser performed an experiment analogous to the one described here to see whether the results of a skewed input would hold in a non-linguistic forced choice task as well. The results indicate that the learning advantage of skewed frequencies is not specific to language.

The experimental evidence discussed here suggests that the frequent use of one verb in a pattern facilitates the learning of the semantics of that pattern. Goldberg's corpus findings demonstrate that this kind of token distribution is available to learners in the input. What this means, is that after hearing *put*

used many times in the VOL caused-motion construction (9), children come to associate the meaning of *put* with the construction's functional semantics regardless of what verb is used (10):

(9) She put a finger on that.

(10) He done boots on. (STE, 28 months; Bates, Bretherton, and Snyder 1988)

(Goldberg 2006: 88)

This data illustrates how the rough constructional meaning of "X causes Y to move Z_{loc}" comes to be correlated with the Subj V Obj $Obl_{path/loc}$ (i.e. VOL caused-motion) formal pattern.

Goldberg clarifies this phenomenon by stating that a single high frequency exemplar facilitates learning but is not altogether necessary for successful acquisition to take place. The relation between form and meaning can still be recognized and cognitively anchored by a form meaning association across several distinct verbs, each with a relatively low frequency. This is evident due to the fact that the balanced condition significantly outperformed those in the control condition. This is quite important because in naturalistic corpus data, there is not always a single verb whose frequency stands out. This is especially the case if constructions are defined as general as possible. The transitive construction does not possess one single high frequency verb but rather a comes to mind.

2.1.1 Motivation for learning Generalizations:

Goldberg cites a number of cognitive psychologists in her explanation of why the generalizations constructions afford are reliably and readily learnt. She indicates that people do not simply use observation or feature learning in order to learn new concepts: they pay attention to the features their prior knowledge mark as important (Murphy 2002: 63; Abbot, Lieven, Tomasello, 2004). Human's ability to categorize is generally motivated by some functional pressure, the fundamental utility of categorization is to be able to predict or infer certain properties of an input on the basis of a perceived set of characteristics (Goldberg 2006: 103). The human cognitive system is not limited to generalization,

nor does it apply this cognitive mechanism randomly. In the case of language, the goal for a learner is to produce and comprehend language: that is for the learner to both understand and to be understood. Goldberg demonstrates that generalizing beyond a particular verb to a more abstract pattern is useful in predicting overall sentence meaning. In fact, knowledge of constructions is in some cases more useful than knowledge of verbs. Others have stressed the importance of weighting different cues, dependent on their reliable co-occurrence with certain outcomes. Goldberg and Casenhiser(2005) hypothesize that predictive value encourages speakers to generalize beyond verbal knowledge to the functional semantic aspect of construction pairings (Goldberg 2006: 105).

For example, it is clear that when *get* appears in the VOL pattern, it signifies caused motion, but when it appears in the VOO pattern it conveys transfer:

a. Pat got the Frisbee over the wall.

b. Pat got Jim a cookie.

Notice then, that *get* in isolation has low cue validity as a predictor of sentence meaning. Since many verbs appear in a variety of constructions with largely different interpretations, speakers ought to attend to the constructions that verbs inhabit. Conversely, if we contrast the semantic contribution of construction and verb with regards to more subtle semantic aspects like manner, the verb proves to be more predictive than the construction. While both *pass* and *mail* can appear in the ditransitive construction, clearly to attain a full interpretation of an utterance, the speaker needs to recognize both the contribution of the verb and the thematic interpretational relations realized by the construction.

Goldberg and Casenhiser (2005) examine the Bates corpus on the CHILDES database in order to examine whether the formal pattern of [Verb Obj Obl_{loc}] predicted the caused-motion semantic function of "X causes Y to move to Z_{loc}". In this study, they calculate the "cue validity" or the conditional probability that an object belongs to a particular category, given that it has a recognizable feature or cue.

P(A|B) is read as the probability of A, given B. The cue validity of VOL as a predictor of caused-motion is thus P("caused motion"|VOL), and in the child directed speech of the Bates corpus, Goldberg reports this probability to be somewhere between .63 and .83. This range depends primarily on how inclusive one takes the notions of caused-motion to be, and likewise how inclusively the VOL formal pattern is defined (Goldberg 2006: 107). Goldberg found that 63% (159/256) of mothers' instances of the VOL structure clearly entail literal caused motion (e.g. put them in the box). Other instances of the VOL formal pattern that did not entail literal caused motion were: future caused motion (e.g. You want them in a cup?), locative adjuncts (e.g. he found the bird in the snow) and verb particle interpretations (e.g. stand it up).

In conclusion, Goldberg has offered two main factors that are likely to motivate speakers to form argument structure generalizations. Initially, children generalize at the level of specific verbs in argument frames because the verb is the best single word predictor of overall sentence meaning. For example, *put* designates something causing something to move, *go* indicates something moving and *make* results in something taking on a new characteristic or becoming something else (Goldberg 2006: 78). In this way children use knowledge of the main verbs found in constructions as a template from which to derive the argument roles of the various constituents in these patterns. Goldberg believes general purpose verbs provide the foundation for both initial semantic and syntactic generalizations, and thus are path breakers on the route to the acquisition of form and meaning correspondences (i.e. constructions).

In Goldberg's forced comprehension task, we see further evidence that the high frequency of prototypical verbs (i.e. a skewed input) optimizes the learning of constructional meaning. Finally, Goldberg shows the motivation for children to generalize beyond specific verbs to form abstract argument structure constructions: constructions have high cue validity as a predictor of overall sentence

meaning. Given the fact that many verbs have low cue validity in isolation (still the best of any single word), attention to the construction's contribution to semantic interpretation is deemed essential. (Goldberg 2006: 126). In the next section, I describe how Anat Ninio critically analyzes a large corpus of speech between children and their parents. To a certain extent, she evaluates whether the corpus does in fact meet the hypotheses developed by Goldberg. In this way Ninio's work has some relevance to the tenants of Construction Grammar. In its own right, Ninio (2011) hypothesizes and argues for a different model of developmental syntax, namely the network theory described earlier.

2.2 *SDIO's* Methodology and Results

Defining the Corpus:

One of the primary goals of Ninio's project is to characterize the syntax of English speaking children, and any distinguishing features of child directed speech. The child language corpus used in this project comes from CHILDES – the Child Language Data Exchange system – which is a public domain database for corpora on first and second language acquisition. The database contains transcripts from many different research projects. Ninio's project targeted naturalistic interaction between parent and child dyads. The upper boundary for child age was three and a half years. There are a number of criteria that were implemented in selecting which projects available in the CHILDES database to include in *SDIO*'s corpus. First, all project participants had to be native English speakers. All children included in the project had to be normally developing and without hearing or speech problems. Only projects involving parent and child dyadic interaction were included. Projects where children interacted with multiple adults or produced monologues were excluded. No transcript was included in the corpus if it did not have at least 100 turns of speech. 33 research projects from the CHILDES archive contained transcripts that met these standards: the British projects, Belfast, Howe, Korman, Manchester, and

Wells, and the American projects Bates, Bernstein-Ratner, Bliss, Bloom 1970 and 1973, Brent, Brown, Clark, Cornell, Demetras, Feldman, Gleason, Harvard Home-School, Higginson, Kuczaj, MacWhinney, McMillan, Morisset, New England, Peters-Wilson, Post, Rollins, Sachs, Suppes, Tardif, Valian, Van Houten, and Warren-Leubecker. (Ninio 2011: 54).

After these project standards were used to narrow down which CHILDES transcripts were to be included, a set of utterance conditions was further selected to filter the data. Only those utterances directly between a parent and child were included (i.e. visitors, investigators, siblings and visitors sentences were excluded.). Only spontaneous speech was kept, reading from books or nursery rhymes was left out of the corpus. Finally, child utterances that were an exact imitation of any of the three previous utterances were left out.

The final corpus included 1.5 million words of parental speech and 200,000 words of child speech. There were 506 different speakers in the parents' corpus and 421 in the child corpus. The threshold for individual contribution to the corpus was set at 3000 sentences. Without necessarily intending this outcome, 93% of the parents speech in the sample talked to their children between the age of 1;0 and 2;6 (Ninio 2011: 55). The corpus was hand-parsed for the three core syntactic relations of the project. Parsing rules were based on principles from the Minimalist Program (Chomsky 1995).

Global Similarities in Grammatical Relation Distributions:

338,970 tokens of the three core syntactic relations were found in the parental corpus, and in the children's corpus a total of 25,769 tokens were coded. There is a striking similarity in the distribution of the three core grammatical relations in the parents' and children's corpus. The relative frequencies of subject-verb, verb-object, and verb-dative object are near identical in the two samples. In the parental corpus 57.6% of the tokens were SV (601 verb types), 40.6% were VO (776 verb types), and 1.8% were VI (66 verb types). The data coded in the child corpus was 55.7% SV (220 verb types), 43.1% VO (238

verb types), and 1.2% VI (24 verb types). (Ninio 2011: 56). The first notable finding from this study is that child speech has a distribution of the clausal relation that is nearly identical to that of their parents. The children manage to recreate the global distribution of the parental corpus with a smaller number of verbs than the parents use. Child speech's clausal distribution is remarkably similar to that of parents, but not because they copy a statistically random selection of parental utterances. Rather, children generate a set of utterances with some differing features, primarily related to the different number of verb types they employee. Regardless, they match the distribution of the parental speech. The corpus also indicates that overall the dative object is very infrequent.

In order to adequately compare the child corpus to the parental corpus, Ninio artificially reduces the 338,970 token adult corpus into 10 random samples with an equal size and distribution to that found in the child corpus. Each of these 10 parental samples contains 14,375 SV tokens, 11,116 VO tokens, and 305 VI tokens from the large parental corpus (Ninio 2011: 64). Consequently, there are three different sample types in total. The child corpus, the 10 reduced sized parental corpora, and the cumulative parental corpus.

With corpora better equipped for comparison, *SDIO*'s focus is still on the clausal core, but now more specifically on the special multiword verbal patterns unique to English. Contemporary English lacks the rich morphology of synthetic languages. Instead, tense, aspect, and modal specifications are represented by auxiliary verbs. These grammatical modulations are operationalized in English by the use of the copula, semi-copula, quasi-auxiliary modal verbs, and dummy verbs for questioning and negation (along with a host of others).

The copula is the verb *be*; the dummy verb mentioned above is *do;* the auxiliary verbs are *be* and *have,* and the modal auxiliaries are *can, may, must, need, ought,* and *have* (Ninio 2011: 106). Ninio makes a distinction between grammatical verbs (the functional verbs listed above that have unique

syntactic purposes) and lexical or content verbs that have pure semantics and clear thematic subcategorization frames (like sing, kick, dance, or fight). In English, all verbs are marked for the person and number of the subject not the object. The subject is positioned preverbally, and is in the nominative case.

The abstract function verbs listed above have unique syntactic behaviors not found in the rest of the verbal lexicon. Grammatical verbs receive complements like lexical verbs, but the difference there is a noticeable discrepancy in the semantic interpretation of the subject. Inflected auxiliary verbs and the copula have no thematic role for the subject. For example, in the sentence '*I am tired'* the copula is the inflected main verb and also the label that undergoes Merge with the subject. This is one of the canonical configurations in which verbs dispense theta roles onto the subject. However, *be* has no thematic role to give. The interpretation and semantic role of the subject is denoted by the adjective '*tired'*. The copular form *am*, other than indicating the present tense, has very transparent abstract semantics. The sentence effectively means that the referent of *I* is part of the set *tired* things. All patterns merging auxiliary verbs and subjects share variations of this divergent behavior where the subject does not receive traditional thematic roles from the verb. In the sentence 'he was swimming fast' and 'he was invited to the meeting', the participles contribute the semantics, but the auxiliaries receive the syntactic subject. (Ninio 2011: 107).

Considering these facts in a developmental light, it may seem that English's unique set of purely grammatical verbs might add a degree of complexity to the learning process. These verbs offer no subcategorization for their subject or object and thus diverge significantly from the use of normal content verbs. It follows from these observations that if the verbs parents use most frequently in the core grammar relations are various auxiliaries, then there may not be a group of legitimate exemplars, from which to generalize the functional semantics of these two word constructions.

Construction grammar treats grammatical relations as meaningful Argument-Structure Constructions which possess a prototypical semantics. The general semantics of patterns are thought to develop from the most frequently modeled verbs. However, Ninio believes that the special features of the purely grammatical verbs as employed in the core grammatical relations are incongruent with the predictions of Goldberg's theories. (Ninio 2011: 112). Under Goldberg's Construction Grammar, the semantics for the subject in the SV relation is 'agent of action'; the prototypical semantics for the direct object constituent of the VO is 'affected object of agent's action', and the prototypical semantics for indirect object in VI is 'recipient of transfer of possession' (Goldberg 1995). Ninio argues that if the most frequent verbs happen not to be actions in SV and VO, or deal with transfer of possession in VI, the constructions responsible for these two word formations will not have corresponding prototypical semantics and thus should not be readily acquired by the children. The most frequent verbs in these clausal structures may not have any lexical semantics at all. They could be the semi-lexical, grammatical, or functional verbs, like those listed above. None of these auxiliaries or modals has pure semantics which could model the argument structures necessary for interpretational generalizations.

It is not clear if this poses an issue for construction grammar. Ninio turns to the data to see whether these characteristics are prominent features of the clausal core in the context of parent child dyads.

2.21 The Three Core Relations, a Closer Look:

In the *SDIO* corpus, the SV grammatical relation in parental speech occurred 195,206 times, and was filled by 601 different verb types. The four most frequent verbs in these SV combinations are the auxiliary verbs *be, do, can,* and *have*. The verb *be* accounts for 52.6% of the total. (Ninio 2011: 114). Although these verbs are the four full auxiliaries in English, *do* and *have* also have content uses. Therefore, Ninio distinguishes between the uses of these verbs that act in the purely grammatical sense

from those that are full content uses. The global distribution of verbs in child speech and parental speech for the SV relation was quite uniform. For parents, 76% (149,235 tokens) of verbs in the SV relation were purely grammatical uses, while 23% (45,971 tokens) exhibited content uses. For the children, 52% (7,523 tokens) were grammatical uses while 48% (6,852 tokens) were content uses. As you can see the great majority of subjects in parental speech are not ones filling some semantic role assigned by a content verb. In child speech *do* and *have* were used in their content sense 37.8% and 68% of the time respectively. In adult speech, the content uses were markedly lower, 5.5% for *do*, and 27.2% for *have*. We can also see from these statistics that child learners do not avoid use of copulas and auxiliary verbs, but they prefer the content uses of *do*, and *have* slightly more than their parents. Regardless, above half of subjects are for auxiliary and modal verbs in both the parental and child corpus (Ninio 2011: 118).

The verb *be* has more than one syntactic use, and therefore its interpretation is multi-faceted. As an aspectual auxiliary *be* is followed by a progressive participle form as in '*Are you pointing at the crowd?*' As a copula, *be* is followed by a predicate complement, which is a noun, adjective, or adverb as in '*He is happy*', and lastly, *be* can serve as a passive auxiliary followed by the past participle in sentences like '*They were invited to the party'*. Because of the similarity between the copula and the passive auxiliary Ninio counts these together as one category. *SDIO* shows that in instances of SV where *be* was the verb, the majority for both child and parent alike was the copula/passive auxiliary use. 87% of the time when a child produced *be* it was a copula/passive auxiliary, while for the adults it was 79%. The proportion of uses that were aspectual auxiliaries was larger in the case of the parents. 22% of parental SV utterances were aspectual auxiliaries while only 12% of child SV utterances were aspectual auxiliaries (Ninio 2011: 114).

2.22 Verb Object (VO) Relation:

English verbs in the VO relation occur in two primary forms. The first is the canonical relation where the verb takes a direct object (e.g. "I ate the hot dog"). Ninio concludes that occurrences of verbs in this type of VO relation do reliably assign a thematic role for their object, namely the (patient or affected object of agent's action) as stipulated by Goldberg. The other VO form is a unique phenomenon widely known as the light-verb construction (LVC). The notion behind the term 'light' is that although verbs in this construction follow the standard verb complement scheme in English, the verbs *take*, *give* etc. do not always realize the full substance of their action semantics. Thus in the clause '*take* a nap, a walk, or a plunge' and '*give* a shout, a pull, or a thought' the verbs act more like licenser of nouns than anything else. One does not actually "take" a "nap" but rather one "naps". However, the verbs in these structures are not entirely without their semantic predicative influence. The difference between 'take a bath' and 'give a bath' is tangible. The noun in the light-verb construction is special, as in all cases it is an eventive noun (i.e. it is a noun carrying the eventive meaning usually represented by the verb.) Ninio argues that verbs in the LVC form of the VO clausal relation do not assign the same semantics as the canonical form. Since the noun is an event rather than an object, the verb's object receives semantics equivalent to 'event participated in, or induced by agent' rather than the expected 'affected object of agent's action' (Ninio 2011: 108). The object term in the LVC form of the VO relation is identical in its coding features to all other object. It is in the accusative case and occupies its normal post-verbal position.

The issue at hand, as before, seems to be distinguishing the real content uses of the verbs in the VO relation, with the notably different semantic role of the object in LVCs. In LVCs the verb is

weakened, the noun is eventive and predicative, and the VO relation has semantics not associated with traditional compositionality. There are no real designations for light-verbs, they exist alongside the normal "heavy" counterparts in the lexicon. *SDIO* opts to let the CHILDES data further inform the conversation on the distribution of the two different VO relations and their semantic differences. (Ninio 2011: 114).

In the adult corpus, parents produced a total of 137,756 tokens of the verb-object (VO) combination, while children produced 11,115. The most frequent verb in both cases was *want*. There were 14,259 tokens of *want* in parents' speech (10.4% of the total) and 2331 tokens (21.0% of the total) in child speech. (Ninio 2011: 119). W*ant* and the other most frequent VO verbs (*have, get, do, give, make, take, hold,* and *put*) reliably assign argument roles to their objects.

SDIO then calculates the proportion of each verb in the VO relation that exists as a LVC. The proportion of *do* and *have* as they appear in the LVC is roughly the same in children and adult speech (i.e. Parents: *do* = 2.86% *have* = 7.8%, Children: *do* = 2.79% *have* = 7.4%). These two verbs have the lowest LVC occurrence rate in both corpora. Children use *make* and *take* in the light-verb construction only slightly less than parents. 16% of VO combinations with these verbs are LVCs for parents, while LVCs make up 12% of the VO combination for children (Ninio 2011: 123). The last verb of interest, *give,* has the highest LVC rate of use in parental speech, with nearly 25% of all instance of *give* meeting the LVC definition. For this verb, children produce many less LVCs than parents. In children's use of the verb *give* only 8% were found to be LVCs.

It is somewhat surprising that children in Stage I of development produce LVCs approximately to the same extent as parents. Including only the five verbs above, 9.5% of all parental VO tokens were LVCs, while that number for children was 6.8%. This is a statistically insignificant difference. Because of this data, Ninio reckons that children seem to be more or less matching their parents' rate of LVC use.

2.23 Verb Indirect Object Relation:

The English syntactic behavior in the verb indirect-object relation (VI) (i.e. that is only indirect objects adjacent to the verb) is related to the analysis of VO in that the VI relation also has instances of the LVC. The verb *tell* was the most frequent occupant of the VI combination in parental speech. Parents produced a total of 6008 VI tokens, *tell* accounted for 1670 tokens and therefore 27.8% of the total VI utterances. The next most frequent verb was *give* and that occurred 1539 (25.6%) times. According to Ninio, t*ell* possesses a thematic role for its indirect object roughly equating to 'addressee of a communication,' Ninio attributes this semantics quality to the indirect object of other VI verbs like *show* and *ask*. (Ninio 2011: 125). The semantics considered prototypical of indirect objects as defined by Goldberg is namely 'recipient of transfer of possession'. This semantic quality was represented in VI utterances by the frequent verbs like *give*, *get*, and *bring*. These three together accounted for 33% of adult VI tokens. As seen in the analysis of parental VO combinations, *give* generated many light-verb constructions. Ninio argues that *give* as it appears in a LVC is not an accurate representative of the semantics 'recipient of transfer of possession'. Examples of sentences found in the VI relation that operate as LVCs include the ditransitive utterances '*give you a bath*' or '*give him a ride'*. Ninio posits that eventive nouns like *a bath* and *a ride* do not refer to objects but events, and therefore cannot be transferred from one possessor to another. With this differentiation established, she estimates that the share of 'recipient of transfer' semantics in the parental VI is closer to 20% than the 33% listed above. (Ninio 2011: 125).

As we know from our investigation of the child produced VO combination, kids use eventive direct objects much less frequently than their parents. In the VI relation children favor the usage of indirect objects to express a thematic 'recipient' rather than an 'addressee of communication'. With regards to the VI pattern, children do not reproduce the same verb distribution semantics as their parents.

2.24 Semantic Assessment of the Clausal Core:

In conclusion, Ninio identifies some discrepancies between the prototypical semantics assigned to each grammatical relation by Construction Grammar, and the semantic notions matching the verbal profiles found in *SDIO's* corpus. In the SV relation 76% of all parental subjects were formal (i.e. not ones that filled a semantic role for the verb.) Parental speech does not shy away from the complex grammatical functions of the SV relation. Thus, children gain heavy exposure to auxiliary verbs and the analytic nature of English syntactic behavior. There was minimal difference in LVC use between child directed speech and adult directed informal speech in the VO clausal relation. The impression Ninio forwards is that parents model, and children are on the way to acquiring, a complete valid English clausal syntax dominated by semi-lexical verbs (e.g. auxiliaries and modals).

This corpus-based study did not find support for Goldberg's (1995, 2006) theory of how construction's prototypical semantics are seeded by a high frequency exemplar. (Ninio 2011: 129). As indicated, the most frequent verb in the SV relation *be* (and the three auxiliaries that follow it in frequency) do not support the 'agent of action' semantics purported by Construction Grammar. The VO constructions most frequent verb in parental speech is *want*. According to Ninio, the direct object of *want* is the 'desired object' not 'the patient or effected object of action.' Finally, the most frequent verb in the VI combination was *tell*. Ninio claims that *tell* has a dative object which designates the 'addressee of a speech act'. She argues that this high frequency occupant of the VI relation does not have the prototypical semantics predicted by Construction Grammar: namely 'recipient of transfer of possession'. Ninio concludes that the core grammatical relations are not meaningful 'Argument Structure Constructions' but purely formal building blocks of syntactic structure, without a related semantic content of their own. *SDIO* thus concludes that children are quite capable of learning formal

grammatical relations that do not possess a pure semantic connection between the verb and the complement. However she does note that other constructions, likely those containing more elements than the word-couplets of core grammatical relations, might be meaningful linguistic signs.

The high rate of grammatical uses in the SV combination, and the frequent production of the light-verb construction supports the notion that children are learning syntax in a lexical-specific, item-specific, use-specific, local manner. What this means is that children learn multiple subcategorization frames or grammatical functions for individual verbs. These are learned on an individual basis. Also the semantics derived from verb-complement pairs are computed in a context specific process. Ninio argues, that the semantic information of a predicate word whether it is a verb, adjective, or adverb is not a uniform a priori constant. Instead, every predicate's semantics is crafted to fit the context of the constituents it interacts with. The example she provides is that the adjective 'old' in the combination 'old friend' means something much different than in the combination 'old relative'. The latter refers to a person's age, while the former refers to the duration of a relationship. (Ninio 2011: 131). In the light-verb construction the pattern receives its meaning largely from the direct object complement and not the combination built on the "heavy" version of the verb. *SDIO* indicates that this fact can only be determined during learning by examining the complement direct object and its possible compositional interaction with the verbs multiple lexical entries. Children are not deprived of the ability to utilize this context-dependent composition of semantics.

Many developmental texts state that early syntactic learning is item-specific in the early stages of acquisition. This stage is thought to be followed by a second stage in which there is abstraction, scheme formation, generalization, and categorization, all based on semantic similarity. Ninio promotes the idea that adult English is much more like young children's early state linguistic system than previously thought. She claims that adult's grammatical knowledge is no less fragmented, structurally or lexically

specific than that of children. This sentiment is not new to the literature. Others have proposed that there is no "grand indivisible unity" where one learns all or none of a linguistic system. Instead language is a wide aggregation of particulars with a weak connective gravity. Each learner acquires more or less of these specific linguistic units according to their means and experience. Ninio has gone to quite a length to make the major conclusion that the clausal core of English is not composed of 'meaningful argument-structure constructions' whose constructional function is seeded by a high frequency neutral verbal occupancy. Nevertheless, *SDIO* regards syntactic learning as a very lexical and structure specific process. This view of syntactic knowledge agrees that the developmental processes deals readily with the particular of form meaning pairs.

Early Linguistic Interactions: Distributional Properties of Verbs in Syntactic Patterns

3 Project Methods:

In this corpus linguistics project, we draw from the same CHILDES project as those used in *SDIO*. Goldberg's *Constructions at Work* also draws from some of the same projects we examine. Because of Ninio's intense assessment of the clausal core, we choose to investigate patterns that are relevant, but not identical to the VO, SV, and VI relations. These associated but distinct grammatical notions are transitivity, intransitivity, and ditransitivity. While these grammatical concepts are fundamentally semantic (i.e. transitivity = a verb that takes a direct object which is equivalent to 'affected object of agent's action') we prefer to start with syntactic definitions of phrasal arrangements that are semantics-free. Sinclair and the COBUILD project (Francis, Hunston, and Manning 1996) identify common patterns of words in English relying heavily on collocation, and phrases. It is beyond the scope of this project to collect data and analyze every verb pattern listed in the COBUILD project. Essentially, the structures defined by COBUILD were developed by close inspection of corpora, and

although they are not fundamentally based on semantics, many of the patterns' usage stand in clear relation to the semantically defined constructions of Goldberg (e.g. VOO, caused-motion VL & VOL) and the clausal structures of Ninio (SV, VO, and VI). The selected COBUILD patterns, and those constructions they encapsulate are listed in the next section.

3.1 Verb Argument Construction Inventory:

The COBUILD grammar project (Francis, Hunston, and Manning 1996) uses The Bank of English (a 250 million word contemporary English corpus) to determine regularities about lexical items and the particular grammatical patterns in which they occur. Every formal pattern in the COBUILD project, which we term a verb-argument construction (VAC), has a simple representation. Each constituent of the pattern is placed in the order it appears. Thus, a typical pattern notation in COBUILD would be **V n that** which simply means 'verb followed by a noun group and a that-clause'. Another VAC in COBUILD notation is: **V *over* n**. This example consists of a verb followed by a prepositional phrase consisting of the word ***over***, and a noun group. This type of pattern notation makes no attempt to indicate the functional category of the elements (i.e. Object, Complement, or Adjunct.) To account for the different operations of constituents within these broad patterns, each VAC is listed with its different 'structures'. The structures of a pattern indicate the varying ways elements after the verb behave semantically (i.e. whether a noun group is an Object, a prepositional Object, a Complement, or an Adjunct). For example, the pattern **V n** is listed with three different structures. Structure I: Verb with Complement "He was my friend." Structure II: Verb with Object "The kid broke a glass." Structure III: Verb with Adjunct "Children talk this way." Structure II of the **V n** VAC corresponds to the definition of transitivity used here.

One of the most innovative features of the COBUILD project is the way it identifies the relationships between a pattern, its various structures, and corresponding meaning groups. One of the

structures for the **V n *into* n** pattern 'Verb with Object and Adjunct' is listed as being occupied by multiple distinct, but interrelated verb groups. One is concerned with causing someone or something to have a quality or idea. This group contains verbs like *breathe, drum, hammer, implant, infuse, inject,* and *strike* (e.g. *infuse* jasmine ***into*** the tea.) Another verb group is related to making someone do something. This group is made up of verbs like *bully, force, nag, con, trick, charm, persuade* and *spur* (e.g. *force* the dog ***into*** the cage.) There are several other meaning groups for this structure. In the COBUILD project, a group of verbs is sometimes divided into sub-groups. For this **V n *into* n** structure, the sub-division might go along the lines of making someone do something by using force, by being nice to them, by deceiving them, or conversely, by motivating them. (Francis, Hunston, and Manning 1996) The COBUILD text has detailed indices that relate verb, pattern, and meaning. This type of grammatical description means the COBUILD project can function as a sort of pattern-based thesaurus.

As you can see the VACs classified in the COBUILD project are combinations of syntactic categories and lexical items. The book presents a thorough catalog of all the English verb patterns. The COBUILD project embraces the reliability and objectivity of computer evidence. It subdivides verbs according to their usage in patterns, and patterns are seen to correlate with meaning. That is to say verbs with similar patterns have similar meanings (Francis, Hunston, and Manning 1996). The Bank of English data is inspected with minimal theoretical presuppositions about grammatical structure. The resulting view of language is one in which no definite distinction between lexicon and grammar can be made. The project characterizes lexical items in terms of their distribution in certain grammatical patterns, these syntactic structures in turn occur only with particular groups of verbs. In this way the standard view of lexical items (i.e. conventionalized utterances that must be learned on an individual basis) is extended to grammatical patterns. This lexicographic description is the index of syntactic patterns from which the current study begins.

The semantics of the caused-motion construction as defined by Goldberg is roughly "X causes Y to move Z_{loc}". This discourse function comes to be associated with the Subj V Obj $Obl_{path/loc}$ structural form. (Goldberg 2006) In this construction, the verb is non-stative and the Obl_{loc} is a directional phrase (e.g. They sprayed paint onto the walls. She put a finger on that). The verb locative construction has similar semantics, but no expressed verbal object. The intransitive motion Verb-Locative has the formal pattern (VL: [Subj Verb Obl_{loc}]). (The dog ran *over* the hills. He walked to the park.) The semantics linked to the VL pattern is roughly "X moves to Z_{loc}." The intransitive (SV) subject has semantics equivalent to "agent of action" while the transitive (VO) object roughly represents "affected object". Finally the indirect object in the VI and ditransitive pattern signifies the "recipient of transfer of possession" semantic role. The COBUILD patterns that correspond with these semantically derived constructions are: **V n n** (ditransitive), **V n** (transitive), **V** (intransitive). Note that many of the VL and VOL caused motion constructions contain instances of transitivity and intransitivity. VACS corresponding to VL caused-motion construction for example: **V *prep/adv*, V *about* n, V *at* n, V *in* n, V *off* n, V *on* n, V *out of* n, V *over* n, V *through* n, V *to* n.** VACS corresponding to VOL caused-motion verb construction include: **V n *prep/adv*, V n *about* n, V n *at* n, V n *in* n, V n *off* n, V n *on* n, V n *out of* n, V n *over* n, V n *through* n, V n *to* n.**

3.2 The Corpus:

To get a representative sample of early learners' linguistic environment we look to the CHILDES database. (MacWhinney, 2000) We consider 'early' as anytime between birth and three and a half years. The corpus for this project is composed of the CHILDES project list that met the criteria used in *SDIO*. Like Ninio, we are interested in only native English speakers on a normal developmental path. We are also particularly interested in linguistic data from child parent dyads, as this environment may embody a unique register with distinct global features. The final corpus for this study consists of all the data from

these CHIDLES projects: Belfast, Howe, Korman, Manchester, Wells, Bates, Bernstein-Ratner, Bliss, Bloom 1970 and 1973, Brent, Brown, Clark, Cornell, Demetras, Feldman, Gleason, Harvard Home-School, Higginson, Kuczaj, MacWhinney, McMillan, Morisset, New England, Peters-Wilson, Post, Rollins, Sachs, Suppes, Tardif, Valian, Van Houten, and Warren-Leubecker. (MacWhinney 2000).

The final collection of data in this study is not identical to that in *SDIO*. We did not apply the same line by line utterance constraints as *SDIO*. That is to say, we did not limit the number of contributions from each speaker to 3000 utterances, nor did we remove child speech that was an immediate imitation of the parent. Lines of transcripts that are readings from books or nursery rhymes were left as part of the corpus. This project did not hand code every utterance of the corpus. We instead used the part of speech and dependency tagging provided. In the full corpus, the total number of verb types used by adults was 1878; the total number of verb types used by children was 1240.

3.3 Searching the Corpus for VAC patterns:

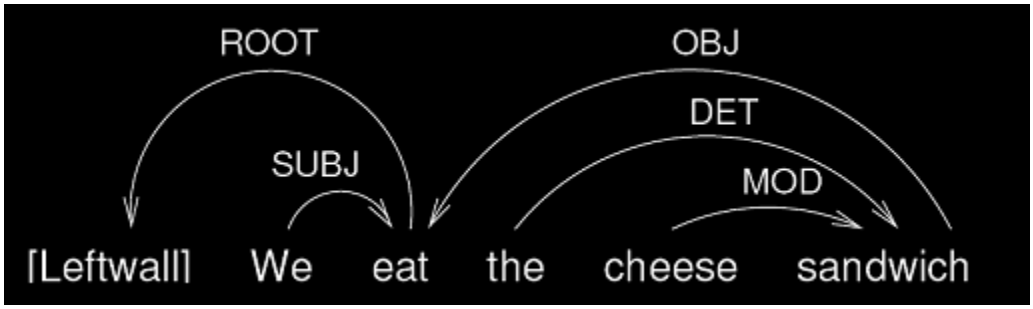
Although The Bank of English used by the COBUILD project has a large spoken component, it is mainly composed of written data. One of our goals in studying a child language corpus is to identify the early lexical specification for VAC patterns. By examining the verb type-token distribution found during parent and child interactions we intend to better describe and understand the relationship between, verb, syntactic form, and communicative function. VACs in the CHILDES corpus were extracted by using search commands in the CLAN environment. The CLAN (Computerized Language ANalysis) (MacWhinney, 2000) environment is a program specifically designed to analyze transcripts in the CHAT format. CHAT is the standardized formatting convention for all transcripts in the CHILDES archive. All of the files are computerized transcripts of recorded speech from developmental studies. The formal specifications of the selected COBUILD VAC patterns are translated into CLAN queries. These queries are designed to retrieve from the corpus instances of the VAC in its designated structural

uses. The resulting data is entered into Excel where we do precision testing and extract the verbal occupant of each identified VAC. The CHAT format used by the CHILDES data allows users to search on different speaker tiers as well as at different levels of grammatical description. Speaker tiers (i.e. *CHI, *MOT, *INV and others) are codes that denote which speaker (child, adult, or interviewer) produced each line of text. This is valuable in that queries can be designated to only inspect certain types of speakers.

Each line of the transcript has a morphological tier (%mor). The difference between the utterance tier and the %mor tier is that the %mor tier is tagged for part of speech. The %mor notation for the word *about* when it is being used as a preposition is thus, prep|about. A full sentence on the %mor line looks like this: pro|I v|think pro|we aux|should v|leave pro:dem|that adv|alone. In conjunction with the COMBO command, we make great use of the %mor tier in finding the caused-motion constructions that are of interest to this study.

The COMBO function provides users with ways to match patterns of letters, words, or groups of words with those in the data files. COMBO uses algebraic symbols to define the words or combinations of words that are to be extracted. In particular there are three symbols to note ^ means ‘immediately followed by’, * when after a part of speech means ‘any lexical item of this category’ and + is the ‘inclusive OR’ connector. In a COMBO command on the %mor tier the user spells out the pattern of interest. Take for example the made up search (v|*^((pro|*+n|*)+(det|*^(pro|*+n|*))^adv:loc|*). This search will return transcript lines that have any verb followed immediately by any pronoun or any noun, followed by any locative adverb. The secondary embedded OR statement (det|*^(pro|*+n|*)), means this search will also return lines with verbs whose pronouns or nouns are preceded by any determiner.

Finally, the English CHAT files come pre-parsed by the GRASP parser. The %gra line represents the dependency grammar relations between constituents of each sentence.



(MacWhinney 2000: 181)

This level of grammatical description was of particular significance to the searches regarding **V n** (transitive), **V** (intransitive), and **V n n** (ditransitive) patterns. The simplest transitive relationship is a verb taking an object (e.g. I saw the boat). An equivalent dependency search would look like this: 1|2|SUBJ^2|0|ROOT^3|2|OBJ. The first number is the position in the sentence (or the serial), and the second number is the pointer. The ROOT always points to the leftwall of the sentence, and in the example above you can see that the object of the verb points to the ROOT (i.e. *sandwich* points to *eat* and *boat* points to *saw*). Using CLAN it is not possible to craft queries specifying that the ROOT be a verb. However, if the utterances are not single word items, the ROOT is generally a verb. The search 1|0|ROOT^2|1|OBJ only looks for sentences where the first word is the verb, followed immediately by the object - something like the fragment "eat cake". In order to catch transitivity relationships that are in non-sentence initial position we transposed the serial and pointers multiple times. For example 1|0|ROOT^2|1|OBJ turns into 2|0|ROOT^3|2|OBJ. There exists a real obstacle in that CLAN does not allow variables as serial or pointer. A more efficient and versatile solution would look like: X|0|ROOT^X^1|X|OBJ. This type of flexibility is available to those using XQuery and the CHILDES data in XML format. Nevertheless, we made the best possible effort to extract a representative proportion of the intransitive and transitive constructions. For the full list of queries used in this project refer to Appendix #1.

Two coders each reviewed 100 random samples of each VAC in order to check the precision of our search methods. The coders inspected each line of transcript returned by the query to see if it met two criteria, (1) does the utterance meet the structural definition of the construction (i.e. Verb Object Obl_{loc})? (2) does the sample utterance meet the semantic interpretation of the particular construction (i.e. X causes Y to move to Z_{loc})? The queries only operate on syntactic information. Therefore this coding allows us to examined how reliably particular classes of syntactic patterns mapp to semantic interpretations. As mentioned by Goldberg there are varying levels of inclusiveness when it comes to coding the semantic content of an utterance. In this project only the strictest sense of literal caused motion was coded as meeting the semantic component of a constructions definition.

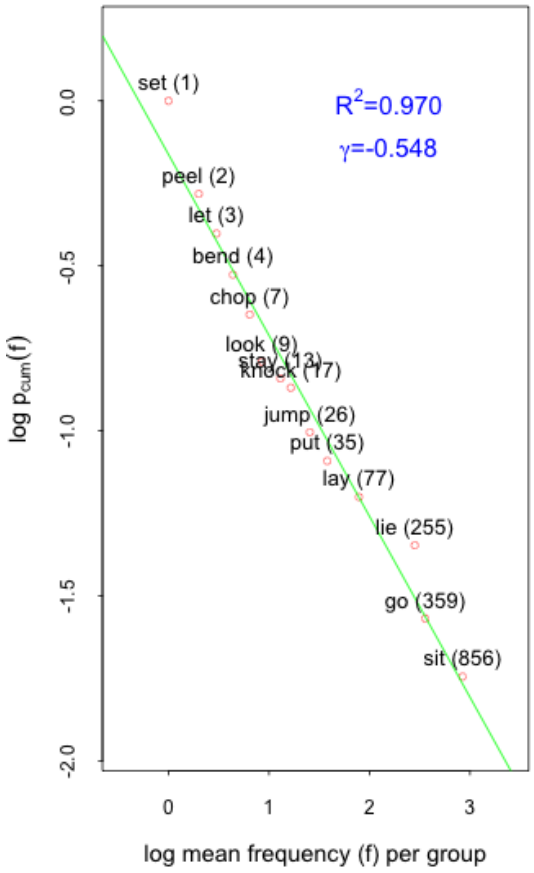
3.4 A Frequency Ranked Type Token VAC Profile:

The queries extract utterances matching the VAC patterns listed above at the rate indicated by our precision testing. We did not check the recall rate of our searches due to time constraints and a lack of man power. Once our search returns the transcript lines matching the queries specifications, the relevant verb was extracted using Excel. In the case of the verb locative and verb object locative constructions, we extracted the first verb preceding the preposition or adverb locative. The resulting lexical information for each VAC is arranged into a verb type-token frequency table (right). Each verb type is accompanied by the number of times that verb appears in the given VAC. Some verbs appear across many different VAC patterns with relative abundance, while other verbs exhibit a selective preference towards

Verb	Constr. Freq	Corpus Freq
sit	856	2689
fall	831	1728
go	359	17622
get	310	17420
lie	255	300
come	80	4925
lay	77	111
slide	41	86
put	35	9803
jump	26	496
run	25	565
want	18	17536
knock	17	443
roll	16	191

appearing in a particular VAC. For example, the verb *scuttle* appears almost solely in the **V across n** VAC.

Zipf's law gives a mathematical description of how a relatively small set of words account for most linguistic tokens in a corpus. In a corpus of natural language, Zipf's law states that the frequency of any utterance is inversely proportional to its rank in the lexical frequency table. (Zipf, 1949) This function means that the most frequent word occurs approximately twice as often as the second most frequent word, and three times as often as the third most frequent word. "If p_f is the proportion of words whose frequency in a given language sample is f, then $p_f \sim f^{-b}$, with $b = 1$" (Ellis 2011: 2). This scaling statistical relation holds across many levels of linguistic representation. The frequencies of phonemes, letter strings, words, grammatical constructs and formulaic phrases all follow this law. (Solé, Murtra, Valverde, & Steels, 2005). The degree to which any data fits a Zipfian distribution is best observed by plotting it on a log-log graph. Switching the axes to a nonlinear scale, a function of the form $y = ax^b$ will appear as a straight line. Therefore we use these logarithmic plots and linear regression techniques to identify the extent to which each VAC verb type-token distribution is Zipfian. The resulting graph for the COBUILD pattern **V *down* n** is shown above and to the right. The VAC type-token list allows us to insect whether the highest frequency tokens in a VAC capture the prototypical discourse function of the pattern (i.e. sit, go, lie, put, jump for **V *down* n**, and get, come, go,



watch, take for **V *out* n**).

3.5 Determining the contingency between construction form (VAC) and function:

Some verbs are closely related to particular constructions. When the contingency between a cue and an outcome is reliable, the more easily the association between the two can be learned. The psychological literature has long recognized that frequency of form is important to associative learning, but cue outcome contingency is the decidedly more important quantity. (Shanks, 1995) Consider the example of a child trying to categorize animal classes. In the set of all *birds*, both eyes and wings appear with the same frequency. However most all animal classes have eyes, while having wings is a highly dependable cue for the outcome "animal categorized as bird". We these adapt these ideas from associative learning theory and posit that constructions with more faithful verb members, and higher cue outcome contingencies should be more readily acquired. Faithfulness is the proportion of verb usage that appear in a particular VAC when compared to the corpus as a whole. Other contingency statistics we use to measure the relationship between VAC and verb are mutual information, and 1-tau^2.

In this study, the 1-tau^2 statistic helps us compute the amount of variance in a VAC verb type-token distribution that is not predicted by overall verb frequency. In this case, tau is the non-parametric regression coefficient between the verb type frequency in a VAC, and that verb type's frequency in the corpus as whole. By taking this value away from 1.0, we are left with the proportion of variance in the verb VAC frequencies that is not explained by the overall corpus verb frequency table. A high value of 1-tau^2 shows that the VAC is heavily selective with regard to verb type. The higher the 1-tau^2 statistic, the more the VAC is attracting particular verbs.

Mutual information is another quantity that measures the dependency between two elements. Given our corpus data, we calculate the individual probability of a particular VAC P(x) and the probability of a particular verb P(y). Then we determine the chance of a verb Y appearing conditioned on a particular VAC pattern X. We also calculate the probability of VAC Y occurring given a particular verb X. These measurements help quantify a VACs degree of lexical specification. If there is high mutual information between verb and construction, than we know certain verbal items have a high probability of appearing in a particular VAC. We see in many cases that VACs indeed have heavy preferences for certain classes of verbs. This is in line with the COBUILD project's methodological design and purpose. The interdependency we observe between VAC and lexical entry is used to help relate verb, grammatical form and discourse function. We see that a great deal of the VACs lexical selection cannot be predicted from inspecting the lexical distribution of the corpus as a whole. Some VACs are much more selective than others.

3.6 Semantic Analysis of VAC verbal occupancy:

There are several ways in which we analyze the semantics of the resulting verb distributions. The semantic measurements we employ are not defined by linguistic distributional information, because that is the dimension of information we have extracted from the corpus. Therefore we avoid latent semantic analysis in favor of other options. The semantic analysis here uses WordNet (Miller, 2009). WordNet groups words into a hierarchical network of synonym groups called synsets. The database is a sort of dictionary and thesaurus. It was also designed to support automatic text analysis and machine learning applications. Every synset contains a group of synonymous words, and most synsets are connected to other synsets via semantic relations like (hypernym, hyponyms, entailment, and troponym). Y is a hypernym of X if every X is a kind of Y (book is a hypernym of novel). The verb Y is entailed by X if

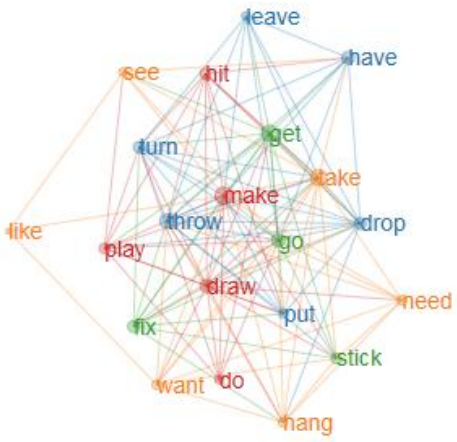
doing X you must be doing Y (to move is entailed by to run). The verb Y is a troponym of the verb X if the activity of Y is doing X in some manner (to whisper is a troponym of to speak).

At the top level verbs are organized into basic types corresponding to the broad pragmatic function of each class (like *move*1 which contains all verbs that express translational movement (i.e. fall, drop), and *move*2 which contains all verbs relating to movement without displacement, (i.e. communication). There are various algorithms available to determine the semantic similarity between synsets. They take into account the distance between conceptual categories as well as the structural relation between constituents in the WordNet structure. (Pedersen, Patwardhan, & Michelizzi 2004). We count the number of verb communities (i.e. top level synsets) that are needed to support the verb networks structure. Each VAC's verb token distribution is fed into WordNet. Frequency does not affect status in the network. Any verb that occurred in a VAC once is represented by a node in the network. However, to clear out parser noise, and insignificant membership, we also calculate network relations for only those verbs that occur ten or more times in a VAC.

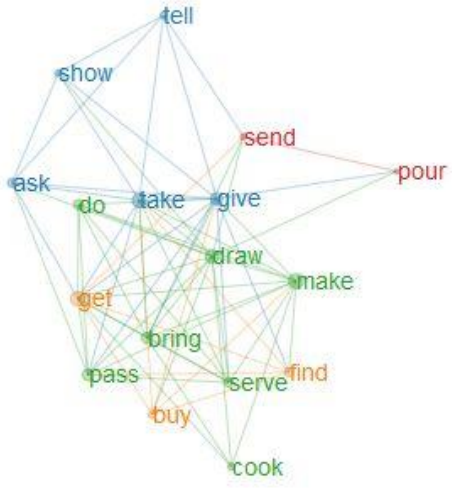
There are a number of network statistics that are observed for each VAC: number of nodes, number of edges (degree), network density, average clustering, and modularity. Network density is defined as the ratio of existing edges to potential edges. Therefore if a network has a density of 1 every node in the network is connected to every other node. Network density is a statistic that helps indicate the interconnectedness of a node. Average clustering is another measurement of how nodes in a graph tend to cluster together. If node A is connected to three other nodes (B, C, D), and each of those three nodes is connected to each other (B-C, B-D, C-D) than the clustering coefficient = 1. If node A is connected to three other nodes (B, C, D) and there is two edges between the three (e.g. B-C and B-D but *not* C-D) than the local clustering coefficient of node A = 2/3.

A hub is defined as a node in a network with a very high degree (i.e. number of connections or edges). Hubs serve as central connections in a network and they often mediate shorter path lengths between many disparate nodes. Hubs are an essential feature of small-world networks. In a graph of a small-world network, most nodes are not neighbors (connected) to one another, but most nodes can be reached from every other node by a small number of steps. As such, small-world networks have low mean-shortest path length between nodes. If a network has a degree-distribution which can be fit with a power law (i.e. a few nodes account for a large proportion of the total network degree) that network is said to be a small-world network. Hubs are the center of networks, and in the case of WordNet, a verb being a hub is an index of that verb's prototypicality.

Modularity is designed to measure the strength of division of a network into modules, also called clusters or communities. Networks with high modularity have dense connections between nodes within modules, but sparse connections between nodes in different module communities.



CHI VLon Semantic Network



CDS Ditransitive Semantic Network

In the CDS ditransitive example above, *give* and *get* look to have particularly high degree and are therefore noteworthy hubs.

4 Results:

Our core research question is concerned with to what extent VAC form, function and usage promotes robust learning of five classes of VACs: **V n n** (ditransitive), **V n** (transitive), **V** (intransitive), **V prep/adv** (VL caused-motion), and **V n prep/adv** (VOL caused motion). Drawing from Goldberg (2006), and empirical work in the psychology of learning, we hypothesize that learnability is optimized for constructions when they are (1) Zipfian in their type-token distribution, (2) selective in their verb occupancy, (3) semantically coherent in their verb occupancy, and (4) have high frequency verb types that serve as prototypical exemplars for the generalization of the syntactic pattern's function. (Goldberg 2006, Ellis 2011: 6). The methods outlined above enable us to examine the degree to which natural speech between children and parents meet these associative learning criteria.

In the CHILDES data the Obl_{loc} element of the two caused-motion constructions is tagged as one of two parts of speech, either preposition or adverb locative. The queries reflected this fact, and so independent precision testing was done on the two Obl_{loc} varieties of the VL & VOL VAC. The mean precision for the VOL_{advloc} VACs is 87% and the mean precision for VL_{advloc} VACs is 72%. The precision for individual VACs in these construction classes can be found in the appendix #2. [1] What the precision statistic indicates is that 87% of the utterances (in our random sample) returned by the VOL_{advloc} query, and 72% of those returned by the VL_{advloc} query match both their construction's structural definition and semantic utility.

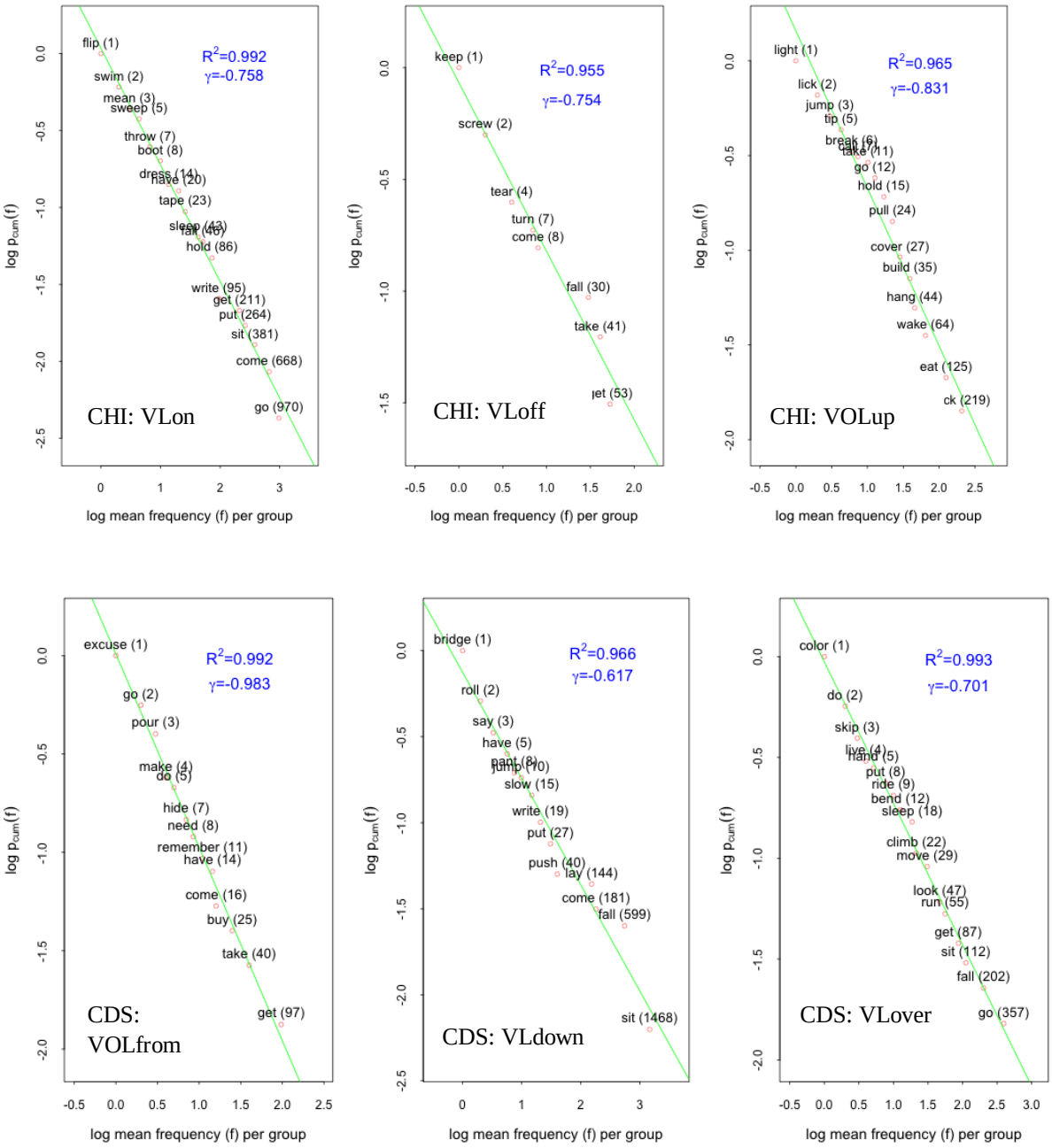
[1] Currently the final the precision values for each VAC are still being calculated. They will be completed shortly after this submission. Cue Validity will be added to the report and appendix.

4.1 Zipfian type-token usage distribution & Prototypical Exemplars:

Our research demonstrates that all VACs in the parental and child directed speech are Zipfian in their verb type-token usage. In the evaluation of our verb token distribution plots, the gamma value (γ) represents the slope of the line of best fit, and the regression value (R^2) is the proportion of variability in the data set explained by Zipf's power law. Below is a table containing the R^2 and γ value for the five construction classes.

Degree of Variance Explained by Zipf's Law					
Child Directed Speech			Child Speech Production		
	R^2	γ		R^2	γ
Mean VL	0.9661	-0.6074	Mean VL	0.9466	-0.7068
Mean VOL	0.9742	-0.7309	Mean VOL	0.9673	-0.7835
Intransitive	0.994	-0.714	Intransitive	0.987	-0.803
Ditransitive	0.936	-0.461	Ditransitive	0.951	-0.56
Transitive	0.976	-0.621	Transitive	0.976	-0.62

Manual inspection of the VAC graphs below show that the highest frequency verb types consistently take the lion's share of the verb token distribution.



As found in previous research (Goldberg 2006; Abbot-Smith, Lieven, Tomasello, 2004), the handful of verb types that account for large proportions of the VL & VOL token distributions match the prototypical semantic function of the constructions they occupy. Here we look at two individual VOL

caused-motion VACs and two VL caused-motion VACs. One of each is from CDS and the other is from child speech.

Let us more closely investigate the **V *over* n** VAC **(**i.e. the Verb Locative construction with Obl_{loc} constituent *over*) as it appears in child directed speech. The token distribution for this VAC is highly Zipfian with an R^2 value of .993. Our query for this VAC returned a total of 1921 utterances whose verb slot was filled by 132 different verb types. The top five verbs in this VAC and their distributional proportion are as follows: *go* (18.5%), *fall* (10.5%), *sit* (5.8%), *get* (4.5%) and *run* (2.8%). The verb *go* is a very appropriate prototypical exemplar for the semantics of the VL caused-motion construction "X moves to Y_{loc}". The four verbs following in frequency are also directly related to notions of translational movement. In this VAC, each high frequency verbs serves as a neutral action semantic representative for constructional utility.

VAC:	Highest Frequency Verb	Degree of listed verbs	Total Nodes
CDS: V *over* n	go (18.5%)	61	97

The verb *go* serves as a central hub in the semantic network of this VAC. It has connections to 61 nodes out of a total possible 97. In this verb network, *go* has the highest degree of any node. This signifies that the verb is very prototypical. *Go* is also one of the closest verbs to the highest (i.e. most general) WordNet synset move#1.

In child speech, the **V *through* n** VAC (a VL caused-motion construction) has a R^2 value of .911. The VAC appears in 126 utterances with 25 verb type occupants. The five most frequent verbs and their distributional proportions are: *go* (41.9%), *get* (14.1%), *come* (14.1%), *look* (9.2%), *and see* (4.9%). Once again, *go* dominates the token distribution and is an appropriate prototypical exemplar. *Get* and *come*, are also aptly suited to seed the constructional semantics of the VL caused-motion construction.

VAC:	Highest Frequency Verbs	Degree of listed Verb	Total # of Network Nodes
CHI: V *through* n	go (41.9%)	go (15)	24

Go is identifiable as a central hub in this child speech VACs semantic network. It is the second most connected node, and connects to well over half of the other semantic net constituents.

In child directed speech, the token distribution for the **V n *off* n** VAC (a VOL caused-motion construction) has an $R^2 = .961$. This VAC appeared in 369 utterances with a 53 different verbs. The five most frequent verbs and their distributional proportions are: *take* (40.3%), *get* (13.8%), *pull* (5.1%), *turn* (4.0%), and *eat* (2.7%). While this VAC meets the structural criteria of the VOL form namely: [Verb Obj Obl_{loc}] the most frequent verbs in this set must be interpreted carefully with regards to the stipulated constructional semantics of "X causes Y to move to Z_{loc}". Consider this example from the CDS corpus:

"Don't take the tape off the rug." New England corpus data 04.cha *Mother*
"You get the paint off the counter." Valian corpus data 09c.cha *Mother*

In both the use of *take* and *get* in these **V n *off* n** examples, the object Y (i.e. the tape, the paint) is not being moved to location Z (i.e. the rug, the counter), instead the location Z is where the object Y currently resides. In this way location Z is not specifying the destination of movement, but rather the launching site of movement. The interpretation of these sentences in this manner means *take*, and *get* may possess a somewhat different constructional function than that of the VOL construction. The lexical semantics of the preposition *off* is an import part of this interpretive distinction, and may be the cause for this divergent reading. In reality there is not a great deal of difference between "X causes Y to move from Z_{loc}" and "X causes Y to move to Z_{loc}". Both very much deal with an agent, causing an object to move, aspect of interpretation regarding the location shows variation. It is not clear if both of these semantic functions can be considered the outcome (function) of a single cue (syntactic pattern). We expect examples like this to lower the overall cue validity of this VAC pattern.

In child speech, the **V n *on* n** VAC type-token distribution has a regression value of $R^2 = .977$. This VAC occurred in 2602 utterances, and was occupied by 132 verb types. The two most frequent verbs and their distributional proportions are: put (56.9%), and get (7.9%). As found in Goldberg's own corpus analysis, the verb *put* is extremely frequent in both child speech and CDS occurrences of the VOL caused-motion constructions. Some child produced examples of the **V n *on* n** include:

"Put it on the bath." Feldman corpus data 12d.cha *Steven* age: 2;1.

"I got it on my shirt." Kuczaj corpus data abe054.cha *Abe* age: 2;11.

As demonstrated by these examples, both *put* and *get* in the **V n *on* n** VAC provide strong verbal exemplars from which the VOL caused-motion pattern's semantic utility can be generalized.

VAC: VOL Construction	Three Highest Frequency Verbs	Three highest Degree Verbs	Degree of Frequent Verb Not in Top 3 Degree	Total # of Network Nodes
CHI: V n on n	put (56.9%), get (7.9%), have (5.5%)	take (62), get (53), change (52)	put (32), have (34)	96

Put and *get* are shown to have generic action semantics and high connectivity in the WordNet for this VAC pattern. *Get* has the second highest degree of any node, and *put* has 1/3 of all possible connections.

All of the VACs in this study can be classified as either meeting the structural criteria of either the VL or VOL caused-motion construction. Now that we have investigated a few specific VACs from each of these categories, we look to the distributional behavior of the these broader construction classes.

4.11 Distributional Characteristics of pooled VACs in the VL & VOL classes:

The VL ([Subj Verb Obl_{loc}]) and VOL ([Subj Obj Verb Obl_{loc}]) caused-motion constructions relate to a number of COBUILD VAC patterns. Above, we reported a handful of specific VACs and the particular properties of their verb type-token profiles. The values for every individual VAC in this study can be found in the appendix #4. In the following section we pool the data from every VAC into the broader categories of verb locative and verb object locative caused-motion constructions.

In child speech, the cumulative VL caused-motion construction featured 620 verb types and a total of 28650 utterances. The top five verb types and their proportional frequency in this VAC are *go* (23.6%), *come* (9.2%), *get* (8.4%), *look* (7.5%), and *sit* (6.8%). In child speech, the five most frequent verbs in the VL caused-motion construction VAC class serve as worthy exemplars for this construction's prototypical semantics "X moves to Y_{loc}". Some VL caused-motion examples from the child language corpus include:

"Go into the tunnel." New England Corpus data 02.cha *Margaret* age: 2;8,
*"*Come to Child's house!" Valian Corpus data 03a.cha *Child Speaker* age: 2;1,
"I get on you." Valian Corpus data 16a.cha *Child Speaker* age: 2;5.
"Look at Eve." Brown Corpus data eve09.cha *Eve* age: 1;10,
"I sit on it." Valian Corpus data 04b.cha *Child Speaker* age: 1;9.

It is clear that these five verbs are prototypical and lack highly detailed action semantics. [We do not currently have semantic networks for the pooled VAC data.] One may argue that *look* does not entail translational movement, and rather involves the directing of one's visual perspective (e.g. Look out the window.) The example listed here would not meet the functional criteria during coding. This does not mean "move" out the window, instead it means guide your visual attention to the location outside the window. In many cases a person has to move to a vantage point that allows them to see out of the window, but this location is notably different than the area their view is being directed towards. In the discourse context of "Where is my toy?" the response "Look on the bench" matches the X moves to Y_{loc}. One must move to the bench in order to look on its surface.

In the pooled VAC data encompassing all of the VOL caused-motion constructions children used 428 different verbs in a total of 16494 utterances. The top five verbs in this construction are *put* (27%), *get* (11.5%), *have* (5.5%), *take* (4.6%), and *want* (4.5%). The verbs *put*, *get*, and *take* certainly match the

prototypical semantics of the VOL caused-motion construction (i.e. X causes Y to move Z_{loc}). Some VOL caused-motion examples from the child language corpus include:

"Do you put it on the floor?" Brown Corpus data adam33.cha *Adam* age: 3;6
"I got butter on my chin." Manchester Corpus data becky23b.cha *Becky* age: 2;7
"Mom can I have marshmallow in it?" Brown Corpus data sarah086.cha *Sarah* age: 3;11
"You took it outside?" Valian Corpus data 12b.cha *Child* age: 2;6
"Want mine in there." Manchester Corpus data anne11b.cha *Anne* age: 2;1

It is also very possible for instances of *want* and *have* to appear as structural members for this pattern's functional semantics. (I have milk in the fridge. I want the toy by her head. I want some soap in the bath.) However, both *have* and *want* are known to exhibit auxiliary grammatical uses alongside the content uses exemplified above. *Want* is a semi-auxiliary verb, and the *want* acts as an incipient modal auxiliary in its form, *want to* or *wanna* (e.g. I want to go to the store.) (Ninio 2011: 119). An example of *have* in its grammatical form would be "I have got toys" or "I have to put this in the jeep." When *want* or *have* are utilized in these non-content senses, they do not behave as effective prototypical exemplars of the construction. We have not calculated what proportion of *want* and *have* are in the content form. In determining the cue validity of both *want* and *have* in this construction a great deal of variability would be left up to inclusiveness of metaphorical or indirect aspects of motion. In the sentence "I have milk in the fridge" the Obj *milk* is not being moved to the Obl_{loc} rather *the milk* is in the possession of the person *I* and its current location is *the fridge*.

In *SDIO* Ninio argues that the verb *want* in its content sense assigns its object with the thematic properties roughly equivalent to "object of desire." She would claim that *want* does not serve as a prototypical exemplar for the VOL caused-motion construction. However Ninio's interpretation overlooks an important aspect of *wanting* an object in the physical world. When a person desires an object, it necessarily entails that said object is not immediately accessible to them. This means that

something or someone must cause it to relocate. In this way *want* is somewhat related to the pragmatic function of the VOL caused-motion construction. Again we do not mean to dismiss Ninio's strong claim. In the sentence "I want a cookie in my mouth table." the *cookie* is not moving to the *table*, its current location is on the surface of the *table*. In conclusion, w*ant* and *have* do not lend themselves to be an effective exemplar from which to generalize translational movement. This analysis lessons the most conservative estimate of the cue validity for this VAC.

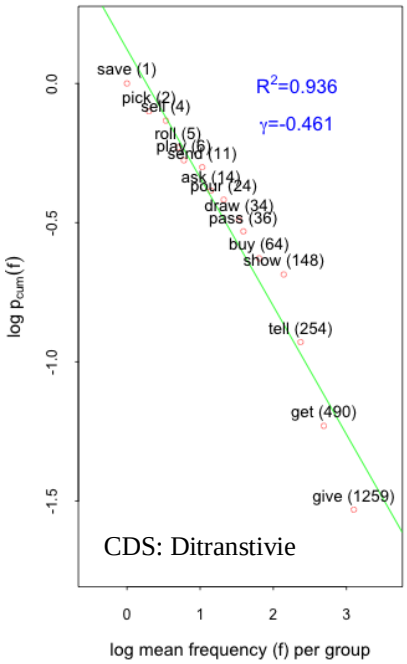
In the CDS, the pooled VACs in the VL caused-motion class occurred 72294 times, with a total of 1033 verb types. The top five most frequent verbs and their proportional frequencies are: come (16.6%), go (15.8%), look (10.8%), get (5.9%), and sit (5.6%). Again at least four of the highly frequent verbs in the parental VL caused-motion construction promote the semantic notion of X moves to Y_{loc}. Alone, these four verbs account for over than 40% of the utterances classified as VL caused-motion constructions.

In the CDS corpus, the cumulative VOL caused-motion construction occurred 12591 times, with a total of 731 unique verb types. The top five most frequent verbs and their proportional frequencies are: put (26.8%), get (7.6%), have (6.9%), take (5.0%), and see (3.1%). The verbs *put*, *get*, and *take* together account for 39.5% of all VOL instances in parental speech. These three verbs are appropriate paradigms for the VOL semantic function X causes Y to move Z_{loc}. *See* possesses many characteristics like the verb *look*. S*ee* does not fit the pragmatic utility of the VOL caused-motion, it does not relate directly to translational movement of an object.

Overall, the highest frequency verb tokens in both the child speech and CDS pooled caused-motion constructions have strong affinities to the hypothesized function of the construction. The four specific VACs and two constructions classes covered in this section meet our fourth criteria for optimized construction learnability: high frequency prototypical exemplars.

4.12 Ditransitive Construction (VI):

Our ditransitive search returned 4506 utterances from the CDS corpus and 3642 utterances from the child speech corpus. This was a small enough dataset to be coded entirely by hand. Therefore, all utterances not matching our **V n n** form and function were excluded from this particular verb type-token distribution. Since we did not calculate the recall of our search, we do not claim that every true ditransitive in the corpus was successfully extracted. Regardless, we believe this sample distribution is a fair representation of the corpus's contents.



After coding the search of the child directed speech there were 2946 confirmed ditransitive utterances filled by a total of 34 different verbs. The verb token distribution graph for the CDS ditransitive construction has a regression value R^2 of .936 and a gamma (γ) value of -0.461. The top five verbs in this construction's distribution are *give* (42.7%), *get* (16.6%), *tell* (8.6%), *make* (7.3%), and *show* (5%). Some ditransitive examples from the CDS corpus include:

"Can you give me a hug?" Valian Corpus data 01a.cha *Mother* line 215
"I should get you some medicine." Valian Corpus data 03b.cha *Mother* line 3493

Together *give*, and *get* account for 59% of ditransitive verbal occupancy. These two verbs fit nicely with both the 'recipient of transfer of possession' semantics proposed to be the thematic role for this patterns indirect object. There is some disagreement about whether the other verbs are suitable exemplars, including *give* as it appears in Light-Verb constructions. In this study we do not code the proportion for each verb in the ditransitive construction that could be classified as a LVC.

VAC:	Three Highest Frequency Verbs	Three highest Degree Verbs	Total # of Network Nodes
CDS: Ditransitive	*give* (42.7%), *get* (16.6%), *tell* (8.6%)	give (13), get (13), draw (13)	17

As noted in the methods, the verbs in the clausal relations were only included as members in the semantic network if they occurred more than n=10 times. This explains the smaller total number of network nodes. Both *give* and *get*, are highly prototypical and both have a degree of 13 (out of a possible 17).

We must be careful to note the differences between Ninio's coding of the VI clausal relationship with the ditransitive VAC query used here. For example, the utterance "*Tell me slowly"* meets Ninio's VI clausal relation because the verb *tell* has been merged with the indirect object label *me*. However, this sentence does not meet the criteria for the ditransitive construction (i.e. Verb + indirect object + direct object). Regardless of this difference, four out of the top five most frequent verbs returned by our **V n n** ditransitive query match those reported by Ninio for the VI clausal relation. In the hand-coded results from *SDIO*'s CDS corpus *tell* accounted for (27.8%) of the verbs in verb-dative construction (VI). The next four were *give* (25.62%), *show* (14.48%), *get* (5.89%), and *ask* (4.03%) (Ninio 2011: 124). In our automated search on a less strictly filtered corpus, the verb *give* has much greater ditransitive use than that found in *SDIO's* the VI results. In our search g*et* is also nearly three times as frequent.

In *SDIO*, Ninio promotes an analysis where the thematic role for the indirect object of *tell, show,* and *ask* is "addressee of a communication'. The claim forwarded is that the thematic notion of 'addressee of communication' is quite different than the prototypical semantics as hypothesized by Construction Grammar: 'recipient of transfer of possession'. This analysis is a marginal judgment. For example, in the utterance "Tell me a story", the indirect object *me* can be interpreted not only as

'addressee of communication', but also as the 'intended recipient of information.' When someone is told a story, they receive a reading and depending on their comprehension the informational content therein. I recognize the distinction Ninio embraces in that when someone asks another person to tell them a story, they are denoting themselves as the 'addressee of communication' and the requested communicative transfer has not yet happened. However, when someone says "*John told me the question yesterday.*" They are not merely indicating that they are the addressee of some communication; they are implying that they have received some information, namely – "*the question*." This once more points to the need that not only should VAC verb token distributions be investigated, but each utterance must be hand coded to see if it matches the literal phrasal utility listed by the construction. It is not enough to merely see how a particular verb could function in the desired manner. Each utterance must be inspected by hand to gauge the true cue validity of the VAC.

The concept of possession is somewhat troublesome in that, when someone is *brought*, or *given* a physical object such as an apple or a ball, after transfer, the subject no longer possesses that tangible entity. In the context of communication, when someone is *told* a story, or *shown* a diagram both the subject and the indirect object now possess experience or knowledge of that direct object. Regardless, when someone is told a joke, they are not merely, the 'addressee' of communication, but the 'recipient' of new stimulus.

In *SDIO*'s claim that the VI clausal relation does not exhibit the prototypical semantics described by construction grammar, Ninio appeals to Light-Verb constructions. In sentences like '*give you a bath*' and '*give him a ride*' Ninio promotes the notion that these eventive nouns cannot be transferred from one possessor to another. This reduces the proportion of the VI distribution she lists as meeting the 'recipient of transfer of possession' semantics (i.e. the relations cue validity). These light-verb constructions seem to encompass the notion of 'cause someone to receive experience of, or engage in an

event'. Regardless of whether causing someone to induce or experience an event is indeed transfer of possession, LVCs match the structural definition of the ditransitive construction. The **V n n** VAC is quite Zipfian and by in large, the top verbs act as appropriate exemplars of the 'X causes Y to receive Z' semantics. We do not distinguish between normal uses of *give* or *take* and those which may be deemed LVCs. Determining the exact degree to which the clausal relation VI meets the prototypical semantics of construction grammar depends on which lexical thematic notions are acceptable during coding.

This same discussion relates to the ditransitive results from the child speech register. In child speech the ditransitive construction appeared 758 times with a total of 25 verb occupants. The Zipfian log-log plot of this verb distribution had a regression value $R^2 = .951$, and a gamma value of -.56. The top five verbs included *give* (43.2%), *get* (17%), *make* (8.3%), *tell* (6.9%) and *bring* (3.5%). In *SDIO* children's five most frequent verbs in the VI construction were *give* (45.57%), *show* (12.13%), *tell* (8.2%), *get* (5.57%), and *ask* (5.25%). Again we see that an enormous proportion of the distribution is taken up by *give* and *get*.

VAC:	Three Highest Frequency Verbs	Three highest Degree Verbs	Total # of Network Nodes
CHI: Ditransitive	*give* (43.2%), *get* (17%), *make* (8.3%)	give (9), get (8), make (8)	11

From the information listed here it is also evident that *give* and *get* are hubs in the child speech ditransitive VAC network. This supports the notion that the ditransitive construction is Zipfian in its type-token distribution and contains highly frequent prototypical exemplars.

4.13 Intransitive Construction (SV):

Again before we delve into our report on the **V** (intransitive) construction, we must be careful to note the differences between Ninio's coding of the SV clausal relationship with the intransitive queries used here. For Ninio, the SV clausal relationship is nothing but the formal relationship created by the

merger between a subject noun, and a verb label. Therefore the minimal requirement for an utterance to be coded as SV does not necessarily categorize it as intransitive. This study classifies an utterance as intransitive, if it contains a doable action verb like (go, lie, sit, die, eat, or arrive) that takes no direct object. It becomes quite clear then that both transitive and intransitive utterances fall under Ninio's SV categorization. (e.g. intransitive: *The statistics don't come until Thursday*. transitive: *I can reach the jar*.) In generative syntactic representations of these utterances, the subjects (i.e. *the statistics* and *I*) are merged with the verb labels *do*, and *can*. Therefore, in Ninio's analysis *do*, and *can* are the verbs tallied. It is important to note, that in our work, it is the substance of the entire verb group that is of interest. The auxiliary verbs *can* and *do* modify the action verb of interest. With this regards, we tally the content verb from verb groups like these. It is in this way that the theoretical orientation (i.e. the difference between merge and verb argument-constructions) can alter how verb type-token tables are collected as well as their analysis. Since the function of the verb *be* is that of an aspectual auxiliary, copula, or passive auxiliary, and not a doable action, it does not appear frequently in our intransitive queries results. In *SDIO be* is the most frequent verb found in the SV formal relationship.

In child directed speech the gamma-value and regression value of the intransitive construction verb type-token distribution is $\gamma = -0.714$, $R^2 = 0.994$ respectively. In the child speech those intransitive distributional values are $\gamma = -0.803$ and M $R^2 = 0.987$. It is clear these verb distributions are extremely Zipfian. In the following section we will discuss the degree to which these constructions are selective in their occupancy, and how this impacts the interpretation of this Zipfian distributional characteristic.

In the child speech, our intransitive query returned 9686 utterances with a total of 488 different verb occupants. The top five verbs from the child speech intransitive distribution are: *go* (17.42%), *know* (7.7%), *want* (7.5%), *do* (5.5%), and *let* (4.4%). According to Ninio, the prototypical semantics of the subject in the SV relation is that of 'agent of action' this is also in line with our conception of

intransitivity. The verb *go* and the content use of the verb *do* are worthy exemplars when it comes to assigning this thematic role to the subject.

VAC:	Three Highest Frequency Verbs	Three highest Degree Verbs	Degree of Frequent Verb Not in Top 3 Degree	Total # of Network Nodes
CHI: V or Intransitive	*go* (17.42%), *know* (7.7%), *want* (7.5%)	go (55), run (52), draw (52)	want (25), do (24) know (7)	97

As we have come to expect, *go* is particularly well connected in the semantic network for the child intransitive construction. On the other hand, *want* and *do* have about half the degree.

The verb *know* is quite prolific in the form of "I don't know. I know. How do you know?" and therefore meets the structural criteria of the intransitive construction without representing a doable action. In this way "knowing" is much more related to a mental state rather than a type of action. The subject of the verb *want* represent the 'agent in state of desire' and does not so clearly fit the notion of 'action' either. *Want* ranks highly among our intransitive queries results because the following utterance "I want" meets the structural criteria of the intransitive construction and was matched heavily. Another reason for want's prominence is the incipient modal auxiliary "I want + infinitive verb" which was also heavily matched by our intransitive query. In this structure, the verb *want* does not take an object, but rather an infinitival clause. In this regard it is matched by part of the dependency query for intransitivity, but does not represent a real instance of this grammatical concept.

The verb *let* seems to be prominent in this distribution because the search term (*1|0|ROOT^2|1|OBJ^3|1|COMP^4|1|PUNCT)* was part of the intransitive query. This may have been an inappropriate dependency query to include. *Let* assigns the thematic role of "person permitted to act" to its object. This particular query returns utterances like "Let's see" and "Let's wait." In these instances of *let*, it is functioning as a transitive verb. The complement featured in these utterances (i.e. "we wait,"

and “we see.”) have an air of intransitivity. This moderate distributional proportion of *let* is largely the result of this query being included.

SDIO’s most frequent verbs in the SV relation are all auxiliary function verbs *be*, *do, can,* and *have*. This is due to the very different conceptions of SV and intransitivity. In conclusion, the subject of frequent verbs like *go,* and *do* assumes the ‘agent of action’ thematic role. These verbs take up a significant verbal distribution (~23%). As the same intransitive query used on the child speech was used on the CDS, it too exhibits divergent occupancy.

In CDS, the intransitive query returned 14571 utterances that contained 440 different verbs. We suspect that the number of verb types being lower in CDS than child speech is a result of the fact that any verb that appeared even once (n=1) was included as part of the verbal distribution. If the threshold for admittance was n=10, like in the VAC semantic networks, the CDS verbal repertoire might be larger than the children’s. (Parser noise could also be an issue at hand.) The top five verbs and their proportional frequency in the pattern are as follows: *go* (32.3%), *let* (11%), *know* (9%), *see* (5.7%), and *do* (4.3%). In this distribution nearly one third of all utterances contain the verb *go*.

VAC:	Three Highest Frequency Verbs	Three highest Degree Verbs	Degree of Frequent Verb Not in Top 3 Degree	Total # of Network Nodes
CDS: Intransitive	*go* (32.3%), *let* (11%), *know* (9%),	go (58), work (52), move (50)	tell (12), know (10)	96

The discussion above once applies to the relative strength of some unexpected verbs like *know,* and *let*. Regardless, *go* serves as a strong prototypical exemplar for the SV ‘agent of action’ semantics. Together with the strong exemplar *see* nearly 40% of the distribution is represented by ‘agent of action’ verbs. The problematic dependency arrangement and semantic interpretations remain with regards to *know,* and *let*.

4.14 Transitive Construction (VO):

In child directed speech the slope (γ) value and R^2 value for the **V n** (transitive) VAC are γ = -0.62 and R^2 = 0.976 respectively. In the child speech, the transitive distributions values are γ = -0.62 and R^2 = 0.976 respectively. (There is no error, the values are identical in both speaker distributions.) We can regard these distributions as highly Zipfian in nature.

In child speech, the transitive construction query returned 52,760 utterances that were occupied by a total of 692 verbs. The top five verbs and their distributional proportions are as follows: *want* (10.1%), *get* (10.1%), *put* (7.3%), *let* (5.9%) and *have* (4.8%). Some examples of the **V n** transitive VAC as seen in the child corpus include:

"I want a piece of cheese." Brown Corpus data eve12.cha *Eve* age: 1;11
"I get it." Sachs Corpus data n38.cha *Naomi* age: 2;0
"Put cream on it." Manchester Corpus becky13a.cha *Becky* age: 2;4

Put, have, want and *get* all support the prototypical semantics notion of "effected object of agent's action." As discussed before, there are competing analyses of whether *let,* and *want* fit into this category. We consider "person permitted to act" and "object of desire" to be roughly equivalent to the semantics stipulated for the transitive object. In this perspective, permitting and desiring are doable actions that affect properties of the object.

VAC:	Three Highest Frequency Verbs	Three highest Degree Verbs	Degree of Frequent Verb Not in Top 3 Degree	Total # of Network Nodes
CHI: Transitive	*want* (10.1%), *get* (10.1%), *put* (7.3%)	take (62), draw (57), move (52)	get (49), put (37), want (22)	96

The semantic network for the transitive construction is large and unruly. A very wide range of verbs can take objects and therefore the network is not particularly coherent. While *take* and *move* are the most connected verbs, the frequent verbs relevant to this construction *get* and *put* are still nodes with a high degree of connective. As you can see no one verb takes much more than 10% of the token distribution,

the transitive construction is one case where there is not a single clear prototypical exemplar. It is possible that the group of frequent verbs together forms a coherent exemplar class, but they are not the most flexible or prototypical verb in the semantic network.

In child directed speech, the transitive construction query returned 51177 utterances that occurred with 765 verb types. The top six verbs and their distributional proportions are as follows: *put* (9.6%), *get* (6.7%), *have* (6.3%), *say* (5.9%), *see* (5.5%) and *want* (5.5%). We concluded that the top five verbs in CDS all support the "effected object of agent's action" discourse function of the VO. Many of the previous semantic notions under debate apply to the analysis of the transitive construction as it appears in the CDS. Again we see the lack of a particular exemplar that is both prototypical in its actions semantics and accounts for large swaths of the token distribution. The semantic network for the CDS transitive construction is similar to that of the one in child speech. It is extremely diverse and while some of the high frequency verbs are relatively well connected, they are not the highest degree central hubs.

We conclude that the frequency type-token distributions for the five broad VAC classes in this study are Zipfian. Those R^2 values not immediately listed here can be found in appendix #4. This research also demonstrates that the most frequent verb for each VAC takes the lion's share of the distribution and is prototypical of that construction's functional interpretation, albeit nonspecific in its action semantics.

4.2 Family Membership and Selective Type Occupancy:

In the previous section we saw that each of the VACs verb type-token distribution was highly Zipfian. However, since Zipf's law applies across all levels of language, the nature of these Zipfian distributions is potentially inconsequential. The distributions become more interesting if the verb

form occupancy of a construction is highly selective (i.e. if the frequency of a particular VAC member cannot be predicted from the verbs frequency in the language as a whole.) Using a 1-tau^2 statistic we demonstrate that VACs are selective in their lexical occupancy. The value of this statistic is the amount of variance in each VACs type-token distribution that is unexplained by overall frequencies in the corpus. The mean 1-tau^2 for the five broad VAC classes is listed below. The 1-tau^2 for each individual VAC can be seen in appendix #4:

VAC Verbal Specificity: 1-tau2			
Child Directed Speech		Child Speech Production	
Mean VL	0.86064	Mean VL	0.87559
Mean VOL	0.83248	Mean VOL	0.84583
Intransitive	0.76769	Intransitive	0.77117
Ditransitive	0.70101	Ditransitive	0.61264
Transitive	0.651254316	Transitive	0.53501

The VACs pooled into the VOL & VL categories have verb distributions where over 80% of the variance is unexplained by the corpus lexical frequency. We conclude that the VACs in these construction classes are highly selective in their verbal occupancy. Further inspection of individual VACs reveals that some are much more selective in their verb occupancy than others. We find that VOL and VL constructions whose Obl_{loc} slot is filled with the adverb locative *around* exhibit some of the highest selectivity. (Child VLaround = .989, Child VOLaround = .961). In child directed speech some of the most selective VACs include VLon with a 1-tau^2 of .939, and VOLup with a 1-tau^2 of .925. From these particular examples, and the mean 1-tau^2 statistic values, we determine that the caused-motion VAC classes are particularly selective in the verbal occupancy. This means that these constructions meet the second criteria for optimization of learnability: lexical selectivity. The caused-motion VACs with the lowest 1-tau^2 value included the CDS VOLat construction (.699), the child speech VOLhere construction (.721), and the CDS VLhere construction (.723).

In CDS the 1-tau^2 value for the ditransitive construction is .70 while in child speech 1-tau^2 for the ditransitive construction = .61. Both of these values are low in comparison to the values seen in the pooled VL and VOL caused-motion constructions. The intransitive construction as it appears in the child directed speech has a 1-tau^2 value of .76, while the child speech it has a value of .77. Finally the 1-tau^2 value of the child directed speech transitive construction is .65. For child speech, the transitive construction had a 1-tau^2 value of .51.

These results point towards the conclusion that both the transitive and intransitive constructions are not highly lexically selective. As discussed, these two VACs had much more variability in their verbal occupation, and often the high frequency verbs did seem to serve as indispensible hubs in the semantic networks. Low selectivity indicates a broader and less focus verbal occupancy. This value is congruent with the conclusions drawn from the intransitive and transitive semantic network: these constructions verbal occupancy lack semantic coherence.

Because of these observations we posit that the intransitive and transitive constructions are more multifaceted and abstract than the ditransitive and caused-motion construction. In this sense our corpus evidence agrees with Ninio's findings about SV clausal relation. This is interesting because our analysis focused on verb groups, and not solely the verb immediately preceding the subject. These results point to the fact that the intransitive construction is a meaningless formal relation that is void of any prototypical semantics. The subject verb relation has such a diverse utility that there does not seem to be a robust semantic role assigned to the subject by the verb.

Both Ninio and this study have shown that the semantic role of the object in the transitive relation is reliably "the affected object of agent's action". However this construction is useful in so many domains that there is very little lexical coherence. Too many verbs take too many different objects in too many interpretational contexts for there to be one single or even a handful of critical salient hubs.

Mutual information statistics provide another means of testing contingency between VACs and verbs. We calculate the mutual information between particular constructions and the verbs that appear in them (M MIw-c). We also measure the mutual information for how individual verbs select particular constructions (M MIw-c). Overall in our corpus the pooled VL and VOL VAC classes have higher mean mutual information in the MIc-w relationship. (Child M MIw-c: 10.5, Parental M MIw-c: 9.53, Child M MIc-w: 11.6, Parental M MIc-w: 11.34) This means that when people are given a construction, they have a stronger knowledge of what verbs might occur in it, than if you give them a verb and ask them which construction it is most likely to appear in. Mutual information statistics are available for every individual VAC in appendix #4.

VAC MIwc (verb to construction)			
Child Directed Speech		Child Speech Production	
Mean VOL	9.8686	Ditransitive	10.97478047
Ditransitive	9.6348	Mean VOL	10.9292
Mean VL	9.1616	Mean VL	9.907
Intransitive	5.4638	Intransitive	5.215990275
Transitive	3.956	Transitive	2.538733415

VAC MIcw (construction to verb)			
Child Directed Speech		Child Speech Production	
Ditransitive	12.31400306	Ditransitive	11.99592491
Mean VL	11.7187	Mean VL	11.8003
Mean VOL	11.01	Mean VOL	10.9657
Intransitive	9.335409047	Intransitive	9.344209055
Transitive	9.13361003	Transitive	8.748921798

These mutual information statistics congruent with the 1-tau^2 values discussed above. In the VL and VOL caused-motion and ditransitive constructions, a verb can be a relatively strong predictor of which construction it will occupy. Conversely, the intransitive and transitive constructions have significantly lower MI_{wc} values. There are so many different words that occupy these structures that being given a

particular verb does help much in predicting its occurrence as either transitive or intransitive. The mutual information values are altogether greater in the MI_{cw} values. When it comes to MI_{cw} values, the intransitive and transitive constructions are still less, but relatively speaking, they are closer to the VL, VOL and ditransitive construction values. In conclusion, verbs have a low predictive value for the transitive and intransitive constructions. This is because they are much less lexically selective than the other constructions. Knowledge of the intransitive and transitive syntactic pattern is also less useful in determining verb occupancy.

4.3 Semantic Coherence:

So far we have seen that every VACs verb type-token distribution is highly Zipfian, and therefore meets the first hypothesized principle of optimized learnability. We have examined the semantics of the highest frequency tokens of these VACs and seen in many cases these verbs are appropriate exemplars from which the functional semantics of a construction can be generalized. Lastly, we have seen that some VACs have high levels of lexical specification, particularly those VACs classified in the caused-motion and ditransitive constructions. During this process we used the semantic networks formed from these distributions to gauge the prototypicality of the high frequency verbs by determining if they were hubs.

In order to get a wider cross comparison between the five VAC classes we must look beyond just the highest frequency constituents and examine the lexical network formed by each distribution. In section 4.1 we did preliminary analysis of the high frequent tokens of each VAC distribution. We postulated that nodes with the most edges are hubs and are offer the most general prototypical communicative utility. Let us first consider a different statistic that describes the structure of these lexical networks.

Network density is defined as the ratio of existing edges to potential edges. The mean network density of the child speech VACs categorized as VL caused-motion constructions is .255, and the network density for the child speech VOL caused-motion construction class is: .289. This value indicates a particular characteristic of all the WordNet networks - each node in these child speech verbal networks is connected to approximately one quarter of the total number of nodes in the network. The mean network density of the parental speech VL construction is .238, and the mean network density for the VOL caused-motion construction is .240. In order to understand this measurements significance there needs to be some comparison. Ellis & O'Donnell (2011) first investigated VAC verb type-token distributions in the British National Corpus (BNC). Their searches of this XML parsed corpus were more sophisticated. The mean network density of six VL caused motion VACs (**V across n**, **V about n**, **V against n**, **V in n**, **V off n**, **V over n**) in the BNC is .081.

This simple comparison reveals that the mean network density of the parental and child VACs in the CHILDES corpus is nearly three times as large as those found in the (Ellis O'Donnell 2011) BNC data. 90% of the BNC's 100 million word corpus consists of regional and national newspapers, popular fiction, and academic journal texts etc. The other 10% is made up of transcripts of unscripted informal conversation.

Evaluation of this comparison could lead to a number of conclusions. A primary response is to say that children do not learn the most colorful and highly specific verbal lexicon of a language first. Instead they pick up the simplest and most fundamental items. One potential reason explaining the much lower network density in the BNC VAC data is that it contains many items with the subtle semantic distinctions found in adult writing. In the **V around n** BNC WordNet, there are more nodes with low degree than found in the CHILDES **V around n** WordNet. For example, the child speech and child directed speech verbal networks for the **V around n** VAC do not include verbs like: hover, reverberate, tighten, gaze, echo, sweep, cluster, gather or stroll. These verbs from the BNC data have more distinct action semantics, and as such are specialized and domain specific. This causes them to be connected other nodes inside the lexical network. These types of verbs are not found in child or child

directed speech. They less the networks density due to their sparse connections. It should be noted that prototypical verbs are still present and exhibit high connectivity in BNC semantic network data. The nodes *move,* and *go* in the BNC **V around n** VAC have 53 and 42 edges respectively out of a possible. 98.

The CDS **V around n** VAC network does contain some specialized verbs (i.e. twirl, splash, skate, and wander) with low connectivity, but the network density is high because there are fifteen nodes each with over 35 connections (there being 65 nodes in the network). While a periphery of more detailed verbs exists in CDS, the fundamental verbs at the center of the WordNet are much more highly connected in the CHILDES data than in the BNC data.

This correlates to the notion that children learn verbs that are best suited for satisfying early communicative needs. These verbs are often the least lexically detailed items. These basic verbs have a high number of edges because they relate to so many domains. This is one metric with which to measure prototypicality of lexical items. We conclude that children's linguistic input and output in the VL, VOL, and ditransitive constructions meet our third criteria for the optimization of construction learning: semantic coherence. While the table for each VAC was included in the previous sections discussion more inclusive tables are listed in the appendix 5.

5 Limitations:

We have argued that understanding the distributional organization of language in use is imperative to understanding the acquisition process. However the methodological process used to examine the presence of these features is not without challenges. The most critical step of the methodology has to do with crafting the search queries that retrieve utterances from the corpus. There are a number of obstacles to this process, some of which have already been mentioned. The limitations of the CLAN software suite made broad searches for transitivity and intransitivity difficult. Inaccuracies in part of speech and dependency tagging resulted in a number of false positives. Finally, due to the nature of COMBO (i.e. CLAN's regular expression search method) it was difficult to exclude certain

lines of transcripts, like infinitive clauses and verb particles. All these affected the integrity of the VAC samples.

Based on the results of our precision testing we admit a certain degree of false positives. However, false positives in the VAC samples should reduce the appearance of the organizational principles important to this discussion. Samples with false positives should have less semantic coherence and lexical specification. What we present here is not a full resolution image of the early learner's linguistic environment. We have only started the mapping process, but this initial prospecting has revealed an organization to VAC occupancy that is somewhat distorted, but ultimately unambiguous in its structural properties. In this project much of the labor is carried out by computer applications. In some cases these lists of machine commands responds with finesse and accuracy, and in other aspects they respond clumsily. Ninio and Goldberg avoid a great deal of these concerns by dedicating a significant effort to hand coding the data in every regard. In projects like these, there is a constant balancing act between corpus size, recall, and precision. After working on a million word corpus the idea of an entirely hand coded project is attractive. A smaller selection offers opportunity for more sensitive analyses.

We did not calculate the recall of the queries utilized in this study. In this manner we only know the accuracy of our searches, not how exhaustively the queries returned appropriate matched utterances. In this regard we know the precision of the searches to be acceptable, but we do not know how many relevant utterances were left unexamined.

5.2 Unfinished Corrections

There was a systematic error in 4 of the queries used in this project. In the Combo searches on the morphological tier, vertical bars were left out of the +s pattern definition switch for both the child and parent VL & VOL caused-motion searches. The searches forms as listed in appendix 2 are:

combo +t*MOT +t%mor
+sv*^prep*^((n*+pro*+det*)+((det*+qn*)^(n*+pro*+det*))+((det*+qn*)^adj*^(n*+pro*+det*))) +k
+r2 +u *.cha

The fixed search looks like:

combo +t*MOT +t%mor
+sv|*^prep|*^((n|*+pro|*+det|*)+((det|*+qn|*)^(n|*+pro|*+det|*))+((det|*+qn|*)^adj|*^(n|*+pro|*+det|*)
)) +k +r2 +u *.cha

The difference between the utterances matched between these two searches is this: without the vertical bar, words that matched the first letters of the category labels (i.e. v, prep, n, pro) were returned regardless of whether they actually were the desired part of speech. For example, the search without the vertical bar returns utterances starting with nouns like ‘velvet’ or ‘violin’ because all that was specified was the letter v followed by anything *. The unfixed search did also return every verb in that position because each verb on the morphological line starts with the sequence v|.

6 Discussion and Implications:

The primary findings of this corpus linguistic study confirm that language in use has important properties of distributional organization. Regardless of theoretical positions, these characteristics of the linguistic input are vital aspects of the acquisition environment. The systematicities of English language in use make for robust learnable cue outcome contingencies and functional categories. We have shown for these five construction classes that the frequency distribution of verbs in particular syntactic configurations is highly Zipfian. This research demonstrates that in the majority of cases, one particular verb is much more frequent than the other members of a VACs verbal occupancy. Beside this observation, we confirm the hypothesis that the most frequent verb in each VAC stands as a strong exemplar from which to generalize the functional semantics of the VACs syntactic pattern. There are exceptions to these axioms, found principally in the case of the intransitive and transitive constructions.

As noted by Goldberg, not every syntactic construction has one individual verb that exhibits these features. The transitive construction is a good example of this case. The form meaning correspondence of transitivity might well be learned across several distinct verbs with relatively low frequency. It is important to emphasize that we do not claim the general purpose verbs prototypes are always the first verbs uttered.

This study demonstrates that the VL and VOL caused-motion constructions and the ditransitive construction are quite selective in their verb form family occupancy. Again the transitivity and intransitivity are noteworthy counterexamples. We show that individual verbs select particular constructions and constructions select individual verbs. Generally, for the VACs in this study, there is a greater contingency between construction and verb type than there is between verb and construction type. We also display that the VL, VOL and ditransitive VACs are coherent in their semantics, and that the most frequent verbs are normally central hubs in the VACs lexical networks.

One major conclusion that can be drawn from this study is that syntactic patterns and semantic function of those formal patterns are not to be dissociated. Psychological theory on the statistical learning of categories suggests that the factors promoting learnability discussed here are indeed features influencing the learnability of concepts. As at least three of the broad VAC classes in this study meet the hypothesized requirements for optimal learning, we suggest that these criteria are indeed the input features cognitive mechanisms utilize in learning linguistic constructions reliably. This data supports the notion of usage-based model of language acquisition, where children learn language from the input along with a normal array of genetically endowed cognitive mechanisms. We show that both of the caused-motion constructions and the ditransitive construction meet the hypothesized criteria and are indeed optimized for learnability. These results indicated that Zipfian scale-free type-token distributions unite characteristic semantic functions with characteristic syntactic frames. This is evident in both

language use and language cognition. We believe the robustness of linguistic development emerges from language being a complex system whose dynamics exhibit these psychologically salient organizational properties (Ellis 2011: 6).

The linguistic environment of early learners is conducive to rational contingency learning. As Goldberg's empirical work indicates this type of learning is not specific only to language. Children are intuitive statisticians; in their interaction with the linguistic input, they can successfully weigh the likelihood of interpretation by gathering information about the relative frequencies and cue validity of form-function pairings. Frequency itself is not a sufficient explanation in itself. Semantic basicness, salience, communicative intent, are all major determining factors in the acquisition of language (Ellis 2002).

In this way, it does not seem necessary for children to have a set of predisposed inferences about how the linguistic system works. We have seen that to a great degree, language in use has vibrant regularities that embody all the constraints necessary for acquisition of useful linguistic forms. The linguistic environment occurs in a patterned and probabilistic manner that promotes children to gradually categorize language experience and become entrenched in a system of knowledge: more specifically the language's lexis, phonology, syntax and semantics. (Abbot, Lieven, Tomasello, 2004) Initially VAC patterns are produced with only one or two verbs for a prolonged period. In an exponentially increasing pattern ore verb types came to be used. That is to say there seems to be progressively more verbal use after ten verb types have been used than after five. Children gradually abstract a more flexible and purely syntactic pattern on the basis of early verbs; this growing generalization permits them to use new verbs more and more easily. (Goldberg 2006: 79). Our corpus data drives the notion that language is an emergent complex system.

Language acquisition is a psychological process that emerges from the interaction of cognitive mechanisms and the linguistic environment. The well-studied and immensely complex cognitive processes that affect learning, attention, and association apply to the learner's linguistic input. This interaction compounded with the rich stimulus environments, produce learner's who intrinsically connect meaningless acoustic disturbances (words and sentences) to their referents, thus allowing them to instill meaning into these symbols. Time and time again, we have seen that there are reliable contingencies between syntactic form, and interpretational aspects of construction components.

The competence and performance of human language undoubtedly lies in the plasticity of synaptic connections rather than in tree diagrams, a formal definitions of constructions or theta-theory. But this does not mean theoretical linguistics cannot come to bear on psychological processing of natural language. Do the laws of motion developed by Newton actually exist outside of the textbook? When a person reaches out to catch a falling cup does their brain know the constant defining acceleration under gravity? Where in the world of interacting particles and bodies of mass does one find the formulas of linear momentum written? These formulas are mathematical descriptions of motion. In the known universe these physical interactions are consistent and measurable. The formulas encoding the laws of motion are extraordinarily useful and accurate predictive tools. They stand as the best model to account for these phenomenon. Similarly, linguistics theoretical models cannot be found if one merely slices open a brain. If linguistic theory becomes successful in making the right predictions about the data is it then said to be the knowledge in our minds? In adopting a computational theory of mind, both generative grammar and construction grammar are seen as algorithms, a type of software code, that our neural substrate implements.

We still lack a great unification between theoretical linguistics, language processing and speech production. It is essential to progress that permutations of theoretical adequacy in all of these fields be

explored and empirically tested. Separating linguistic function and form has been one of the most celebrated strategies in linguistic history and has lead to great expansion and progress throughout the field. However, the age of rarifying the anatomy of syntax above of all other characterizations and functions of the human linguistic systems is over. I believe theories of human language that respect the needs, biases, and overarching principles of cognitive systems can provide important insights into how language is acquired, process and ultimately used.

6.2 Future Directions

There are a number of areas that should be pursued further in this research. One obvious continuation of the work done here is to extend the methodology to the remaining COBUILD VAC patterns. A great deal of the evaluations and conclusions made in this study are drawn from a large pooled corpus of child parent interactions. This study lacks intense investigation of individual longitudinal corpora of single children and their particular interface with parental speech. There are many broad an important flavors that we have gleaned from the statistical stew created by this study, but refined manual inspection of longitudinal developmental transcripts may result in a more palatable and delicate set of findings.

There is a need to start analyses on which constructions are learned first, and why. The hypothesis would have to do with what input and pragmatic properties promote these particular constructions to be the earliest form function pairings learnt. An answer would likely be sought by examining in what ways the earliest used constructions are distinct in their suite of, semantic utility, type/token frequency distribution, semantic coherence, or leading verbal exemplars. This is a serious enterprise onto itself, but is undoubtedly worthy of more investigation.

During our precision testing, utterances were coded under two criteria. The first thing that was looked as was whether the matched corporal utterance fit the structural description of the verb argument

construction. For the **V n *in* n** the sentence "Find the cat in the bushes" fits the formal pattern of the verb construction. The second distinction made in the precision coding was whether this utterance met stipulated the semantic function. In the case of this example sentence, the answer is no. The Obl_{loc} element of the VAC in this case is functioning as a locative adjunct rather than the destination of movement. From our precision results we were able to give a rough estimate of each construction's cue validity (i.e. the percentage of utterances that match both form and function). I believe further projects could benefit from the coders receiving explicit definitions of how inclusively utterances were to be marked. However future caused motion, verb particles or locative adjuncts are chosen to be coded, consistency and transparency are absolutely necessary.

These are some methodological issues about the coding semantic properties. But another challenge remains, how can we better define the action semantics of verbs other than their our intuitive knowledge of their purpose and the classifications of dictionaries. These notions work in human interaction, but how does one configure a robotic arm to know the difference between shoving and pressing, or the difference between releasing and dropping? Coding a machine to have distinctive classes of actions like this requires precise delineation of the physical properties of these different kinds of movement. Code like this may come to serve as a better index of action semantic classification than is currently available.

Ellis & O'Donnell (2011) laid the foundation for the methodology used here. In their examination of the BNC corpus they utilized the corpus data in an XML format. This allowed the functional programming language xQuery to be used in the implementation of VAC searches. This more advanced suite in combination with their professional programming ability allowed for a more clean and defined sample of VAC utterances. In a future instantiation of this methodology, choosing the CHILDES data in its XML format along with crafting the VAC searches in xQuery would be advisable.

Appendix:

#1: CLAN Search List

VL caused-motion construction (adverb locative): Child Directed Speech
combo +t*MOT +t%mor +sv*^adv:loc* +k +r2 +u *.cha

VOL caused-motion construction (adverb locative): Child Directed Speech
combo +t*MOT +t%mor
+sv*^(((det*+qn*)^adj*^(n*+pro*+det*+qn*))+(n*+pro*+det*+qn*)+((det*+qn*)^(n*+pro*+det*+qn*
)))^adv:loc* +k +r2 +u *.cha

VL caused-motion construction (preposition): Child Directed Speech
combo +t*MOT +t%mor
+sv*^prep*^((n*+pro*+det*)+((det*+qn*)^(n*+pro*+det*))+((det*+qn*)^adj*^(n*+pro*+det*))) +k
+r2 +u *.cha

VOL caused-motion construction (preposition): Child Directed Speech
combo +t*MOT +t%mor
+sv*^(((det*+qn*)^adj*^(n*+pro*+det*+qn*))+(n*+pro*+det*+qn*)+((det*+qn*)^(n*+pro*+det*)))^p
rep*^((det*^(n*+pro*+det*+qn*))+(n*+pro*+det*)+(adj*^(n*+pro*+det*+qn*)+((det*+qn*)^adj*^(n*
+pro*+det*+qn*))) +k +r2 +u *.cha

Transitive Construction: Child Directed Speech
combo +t%gra +t*MOT +s@transitive5.cut *.cha +k +r2 +u +o%MOR

Contents of transitive5.cut:
1|0|ROOT^2|1|OBJ
1|0|ROOT^2|3|DET^3|1|OBJ
1|0|ROOT^2|4|DET^3|4|MOD^4|1|OBJ
1|2|SUBJ^2|0|ROOT^3|2|OBJ
1|2|SUBJ^2|0|ROOT^3|5|DET^4|5|MOD^5|2|OBJ
1|2|SUBJ^2|0|ROOT^3|4|DET^4|2|OBJ
1|2|SUBJ^2|0|ROOT^3|4|MOD^4|2|OBJ
1|3|SUBJ^2|3|AUX^3|0|ROOT^4|3|OBJ
1|2|SUBJ^2|0|ROOT^3|5|DET^4|5|MOD^5|2|OBJ
1|3|SUBJ^2|3|AUX^3|0|ROOT^4|5|DET^5|3|OBJ
1|4|SUBJ^2|4|AUX^3|2|NEG^4|0|ROOT^5|4|OBJ

Intransitive Construction: Child Directed Speech
combo +t%gra +t*MOT +s@intransitive5.cut *.cha +k +r2 +u +o%MOR

Contents of intransitive5.cut:
1|2|SUBJ^2|0|ROOT^3|2|PUNCT
1|2|SUBJ^2|0|ROOT^3|2|PUNCT

2|3|SUBJ^3|1|ROOT^4|3|PUNCT
1|2|SUBJ^2|0|ROOT^3|2|JCT^4|5|DET^5|3|POBJ^6|2|PUNCT
1|3|JCT^2|3|SUBJ^3|0|ROOT^4|3|PUNCT
1|0|ROOT^2|1|OBJ^3|1|COMP^4|1|PUNCT
1|3|AUX^2|1|NEG^3|0|ROOT^4|3|PUNCT
1|2|SUBJ^2|0|ROOT^3|4|INF^4|2|XCOMP^5|2|PUNCT
1|3|COM^2|3|SUBJ^3|0|ROOT^4|3|PUNCT
1|3|AUX^2|3|SUBJ^3|0|ROOT^4|3|JCT^5|4|POBJ^6|3|PUNCT
1|0|ROOT^2|4|PRED^3|4|SUBJ^4|1|COORD^5|1|PUNCT

Ditransitive Construction: Child Directed Speech
combo +t%mor +t*MOT
+s"((v|ask+v|blow+v|bring+v|buy+v|cook+v|cut+v|do+v|draw+v|drop+v|feed+v|find+v|get+v|give+v|grow+v|hand+v|offer+v|hit+v|kick+v|leave+v|make+v|paint+v|pick+v|play+v|pay+v|promise+v|pour+v|read+v|roll+v|send+v|show+v|spell+v|take+v|tell+v|teach+v|wish+v|write+v|offer+v|allow+v|assure+v|promise+v|afford+v|lend+v|allocate+v|assign+v|cause+v|charge+v|cost+v|cut+v|deal+v|deny+v|design+v|earn+v|feed+v|fine+v|grant+v|guarantee+v|hand+v|keep+v|leave+v|order+v|owe+v|pass+v|permit+v|prescribe+v|promise+v|read+v|refuse+v|save+v|sell+v|serve+v|set+v|spare+v|supply+v|vote+v|him+v|write)^((pro|me+pro|you+ pro|him+ pro|her+ pro|us+
pro|them)+(det*^n*))^((det*^n*)+(qn*^n*)+(n*)+(det*^adj*^n*)+(pro|it+ pro:dem|that+
pro:dem|those+pro:indef|one))" *.cha +k +r2 +u

VL caused-motion construction (adverb locative): Child Speech
combo +t*CHI +t%mor +sv*^adv:loc* +k +r2 +u *.cha

VOL caused-motion construction (adverb locative): Child Speech
combo +t*CHI +t%mor
+sv*^((det*^adj*^(n*+pro*+det*))+(n*+pro*+det*)+(det*^(n*+pro*+det*)))^adv:loc*) +k +r2 +u
*.cha

VL caused-motion construction (preposition): Child Speech
combo +t*CHI +t%mor
+sv*^prep*^((n*+pro*+det*)+((det*+qn*)^(n*+pro*+det*))+((det*+qn*)^adj*^(n*+pro*+det*))) +k
+r2 +u *.cha

VOL caused-motion construction (preposition): Child Speech
combo +t*CHI +t%mor
+sv*^((det*^adj*^(n*+pro*+det*+qn*))+(n*+pro*+det*+qn*)+(det*^(n*+pro*+det*+qn*)))^prep*^((det*^(n*+pro*+det*+qn*))+(n*+pro*+det*)+(adj*^(n*+pro*+det*+qn*)+(det*^adj*^(n*+pro*+det*+qn*))) +k +r2 +u *.cha

Transitive Construction: Child Speech
combo +t%gra +t*CHI +s@transitive5.cut *.cha +k +r2 +u +o%MOR

Contents of transitive5.cut:

1|0|ROOT^2|1|OBJ
1|0|ROOT^2|3|DET^3|1|OBJ
1|0|ROOT^2|4|DET^3|4|MOD^4|1|OBJ
1|2|SUBJ^2|0|ROOT^3|2|OBJ
1|2|SUBJ^2|0|ROOT^3|5|DET^4|5|MOD^5|2|OBJ
1|2|SUBJ^2|0|ROOT^3|4|DET^4|2|OBJ
1|2|SUBJ^2|0|ROOT^3|4|MOD^4|2|OBJ
1|3|SUBJ^2|3|AUX^3|0|ROOT^4|3|OBJ
1|2|SUBJ^2|0|ROOT^3|5|DET^4|5|MOD^5|2|OBJ
1|3|SUBJ^2|3|AUX^3|0|ROOT^4|5|DET^5|3|OBJ
1|4|SUBJ^2|4|AUX^3|2|NEG^4|0|ROOT^5|4|OBJ

Intransitive Construction: Child Speech
combo +t%gra +t*CHI +s@intransitive5.cut *.cha +k +r2 +u +o%MOR

Contents of intransitive5.cut:
1|2|SUBJ^2|0|ROOT^3|2|PUNCT
2|3|SUBJ^3|1|ROOT^4|3|PUNCT
1|2|SUBJ^2|0|ROOT^3|2|JCT^4|5|DET^5|3|POBJ^6|2|PUNCT
1|3|JCT^2|3|SUBJ^3|0|ROOT^4|3|PUNCT
1|0|ROOT^2|1|OBJ^3|1|COMP^4|1|PUNCT
1|3|AUX^2|1|NEG^3|0|ROOT^4|3|PUNCT
1|2|SUBJ^2|0|ROOT^3|4|INF^4|2|XCOMP^5|2|PUNCT
1|3|COM^2|3|SUBJ^3|0|ROOT^4|3|PUNCT
1|3|AUX^2|3|SUBJ^3|0|ROOT^4|3|JCT^5|4|POBJ^6|3|PUNCT
1|0|ROOT^2|4|PRED^3|4|SUBJ^4|1|COORD^5|1|PUNCT

Ditransitive Construction: Child Speech
combo +t%mor +t*CHI
+s"((v|ask+v|blow+v|bring+v|buy+v|cook+v|cut+v|do+v|draw+v|drop+v|feed+v|find+v|get+v|give+v|gro
w+v|hand+v|offer+v|hit+v|kick+v|leave+v|make+v|paint+v|pick+v|play+v|pay+v|promise+v|pour+v|read
+v|roll+v|send+v|show+v|spell+v|take+v|tell+v|teach+v|wish+v|write+v|offer+v|allow+v|assure+v|promi
se+v|afford+v|lend+v|allocate+v|assign+v|cause+v|charge+v|cost+v|cut+v|deal+v|deny+v|design+v|earn+
v|feed+v|fine+v|grant+v|guarantee+v|hand+v|keep+v|leave+v|order+v|owe+v|pass+v|permit+v|prescribe
+v|promise+v|read+v|refuse+v|save+v|sell+v|serve+v|set+v|spare+v|supply+v|vote+v|him+v|write)^((pro
|me+pro|you+ pro|him+ pro|her+ pro|us+
pro|them)+(det*^n*))^((det*^n*)+(qn*^n*)+(n*)+(det*^adj*^n*)+(pro|it+ pro:dem|that+
pro:dem|those+pro:indef|one))" *.cha +k +r2 +u

#2 Available Precision Statistics:

VLadvloc		VOLadvloc	
ADVERB	PRECISION	ADVERB	PRECISION
anywhere	1	about	0.25
around	0.91	after	0.88
at	0.43	at	0.9
back	0.72	before	1
backwards	1	beside	0.5
down	0.8	besides	1
downstairs	1	by	1
everywhere	1	except	0.83
forward	1	for	0.92
forwards	1	from	0.87
here	0.92	in	0.85
in	0.41	into	0.63
inside	0.77	like	0.86
left	0.33	of	0.89
on	0.74	off	0.86
out	0.83	on	0.83
outside	0.95	out	1
over	0.53	over	1
right	1	since	1
there	0.88	to	0.82
through	0.61	under	1
up	0.87	underneath	1
upside	0.94	until	0.75
upstairs	0.96	up	0.9
GRAND MEAN	0.72	with	0.92
		without	0.5
		GRAND MEAN	0.87

#3: VAC log-log type-token frequency graphs & Semantic Networks:

All VAC token distribution graphs and Semantic Networks are available here:
http://141.211.75.204:8111/vactrac

(Note: University of Michigan network access required.)
Images can be delivered by request: liamc@umich.edu

#4 Individual VAC Statistics:

VAC: CDS VL	Types	Tokens	TTR	Types Cover 95% Token	1-tau2	Entropy	M Token*Faith.	Mean MIwc	Mean MIcw	R2Zipf	Gamma Zipf
CDSVLabout	92	601	15.31	62	0.898	4.342	0.316	11.816	12.135	0.941	-0.815
CDSVLaround	67	544	12.32	40	0.857	3.547	0.352	13.323	14.769	0.966	-0.715
CDSVLat	150	5880	2.55	22	0.817	2.111	18.115	6.319	10.116	0.939	-0.637
CDSVLback	79	1573	5.02	25	0.862	2.842	0.559	9.706	11.932	0.92	-0.597
CDSVLdown	159	4258	3.73	34	0.852	3.376	8.708	8.081	12.015	0.966	-0.617
CDSVLhere	131	4627	2.83	15	0.723	2.494	6.288	6.419	9.669	0.962	0.562
CDSVLin	347	7519	4.61	105	0.837	3.986	4.848	6.979	10.445	0.99	-0.704
CDSVLoff	76	581	13.08	47	0.905	3.794	0.167	12.508	13.384	0.944	-0.739
CDSVLon	330	12143	2.72	65	0.940	3.472	22.750	5.903	10.060	0.994	-0.687
CDSVLout	240	3926	6.11	85	0.818	3.981	1.712	8.355	11.295	0.99	-0.69
CDSVLover	132	1921	6.87	56	0.905	3.972	0.779	9.870	12.353	0.993	-0.701
CDSVLthere	174	2195	7.93	76	0.725	4.004	0.533	9.032	11.073	0.974	-0.719
CDSVLthrough	58	454	12.78	36	0.958	3.383	0.154	13.113	13.575	0.941	-0.706
CDSVLto	205	5704	3.59	47	0.887	3.679	7.148	7.255	10.734	0.981	-0.631
CDSVLup	245	5037	4.86	97	0.926	4.456	9.255	8.744	12.226	0.99	-0.715
Mean CDS VL Class	165.667	3797.533	6.954	54.133	0.861	3.563	5.446	9.162	11.719	0.966	-0.607

VAC: CDS Core	Types	Tokens	TTR	Types Cover 95% Token	1-tau2	Entropy	M Token*Faith.	Mean MIwc	Mean MIcw	R2Zipf	Gamma Zipf
CDSDitransitive2	34	2946	1.15	11	0.701	3.006	12.171	9.635	12.314	0.936	-0.461
CDSIntransitive2	440	14571	3.02	105	0.768	3.819	13.133	5.464	9.335	0.994	-0.714
CDSTransitive2	765	51177	1.49	148	0.651	4.720	86.025	3.956	9.134	0.976	-0.621

VAC: CDS VOL	Types	Tokens	TTR	Types Cover 95% Token	1-tau2	Entropy	M Token*Faith.	Mean MIwc	Mean MIcw	R2Zipf	Gamma Zipf
CDSVOLabout	50	524	9.54	24	0.769	2.949	0.444	10.832	9.412	0.942	-0.719
CDSVOLaround	67	384	17.45	48	0.878	3.645	0.237	13.435	13.507	0.955	-0.837
CDSVOLat	105	1077	9.75	55	0.699	3.923	0.119	10.004	10.682	0.989	-0.803
CDSVOLback	95	2248	4.23	27	0.803	2.991	1.088	8.780	11.337	0.953	-0.642
CDSVOLdown	132	1889	6.99	53	0.836	3.791	1.042	9.582	11.831	0.993	-0.717
CDSVOLfrom	75	375	20	57	0.877	4.284	0.065	12.732	11.879	0.992	-0.983
CDSVOLhere	112	975	11.49	68	0.756	4.222	0.121	10.122	10.295	0.993	-0.853
CDSVOLin	292	9124	3.2	72	0.841	3.281	11.681	5.949	9.330	0.984	-0.709
CDSVOLof	209	3491	5.99	75	0.735	4.163	0.618	7.612	9.499	0.97	-0.719
CDSVOLoff	53	369	14.36	35	0.940	3.600	0.094	12.394	11.011	0.961	-0.814
CDSVOLon	213	6045	3.52	61	0.838	3.181	5.673	6.384	8.960	0.989	-0.708
CDSVOLout	163	3688	4.42	59	0.924	3.896	2.867	8.322	11.014	0.976	-0.688
CDSVOLover	144	1495	9.63	75	0.859	4.203	0.992	10.176	11.884	0.98	-0.785
CDSVOLthere	107	952	11.24	60	0.732	3.866	0.179	10.015	10.411	0.993	-0.765
CDSVOLthrough	47	291	16.15	33	0.906	3.657	0.126	13.659	12.644	0.998	-0.809
CDSVOLto	142	3002	4.73	46	0.805	3.984	2.481	8.625	11.469	0.955	-0.067
CDSVOLup	220	3980	5.53	94	0.954	4.647	7.208	9.142	12.006	0.939	-0.807
Mean CDS VOL Class	130.941	2347.588	9.307	55.412	0.832	3.781	2.061	9.869	11.010	0.974	-0.731

VAC: Child VL Class	Types	Tokens	TTR	Types Cover 95% Token	1-tau2	Entropy	M Token*Faith.	Mean MIwc	Mean MIcw	R2Zipf	Gamma Zipf
ChildVLabout	28	146	19.18	21	0.981	3.770	0.637	14.836	14.074	0.896	-0.927
ChildVLaround	36	273	13.19	23	0.989	3.333	0.367	12.595	12.351	0.897	-0.687
ChildVLat	66	1690	3.91	16	0.898	1.956	7.743	7.729	10.681	0.87	-0.594
ChildVLback	50	758	6.6	16	0.844	2.846	0.524	9.713	11.312	0.896	-0.5
ChildVLdown	111	3185	3.49	24	0.918	3.315	9.941	6.676	10.875	0.97	-0.548
ChildVLhere	96	873	11	53	0.848	3.753	0.431	9.971	12.397	0.952	-0.721
ChildVLin	233	4305	5.41	80	0.825	3.575	3.270	6.327	10.127	0.994	-0.726
ChildVLoff	32	187	17.11	23	0.839	3.377	0.065	13.174	12.150	0.955	-0.754
ChildVLon	234	4094	5.72	92	0.918	4.028	3.950	6.907	10.711	0.992	-0.758
ChildVLout	151	2243	6.73	60	0.862	3.650	1.229	7.862	11.383	0.978	-0.669
ChildVLoutside	32	271	11.81	19	0.922	2.664	0.118	12.038	11.528	0.92	-0.668
ChildVLthere	168	931	18.05	122	0.773	4.324	0.535	10.055	12.187	0.987	-0.979
ChildVLover	83	762	10.89	46	0.851	4.092	0.522	10.241	12.071	0.958	-0.765
ChildVLup	156	2624	5.95	64	0.924	4.151	9.998	8.234	12.109	0.995	-0.675
ChildVLthrough	25	162	15.43	17	0.723	3.021	0.044	14.083	13.454	0.911	-0.645
ChildVLto	97	1760	5.51	30	0.895	3.310	1.409	8.072	11.395	0.975	-0.661
Mean Child VL Class	99.875	1516.500	9.999	44.125	0.876	3.448	2.549	9.907	11.800	0.947	-0.705

VAC: Child Core	Types	Tokens	TTR	Types Cover 95% Token	1-tau2	Entropy	M Token*Faith.	Mean MIwc	Mean MIcw	R2Zipf	Gamma Zipf
ChildDitransitive2	25	758	3.3	12	0.613	3.022	3.449	10.975	11.996	0.951	-0.56
ChildIntransitive2	488	9686	5.04	196	0.771	4.510	8.009	5.216	9.344	0.987	-0.803
ChildTransitive2	692	52760	1.31	136	0.535	4.671	200.715	2.539	8.749	0.976	-0.62

VAC: Child VL Class	Types	Tokens	TTR	Types Cover 95% Token	1-tau2	Entropy	M Token*Faith.	Mean MIwc	Mean MIcw	R2Zipf	Gamma Zipf
ChildVOLabout	19	74	25.68	16	0.938	3.223	0.045	14.371	10.643	0.973	-0.848
ChildVOLaround	34	101	33.66	29	0.962	3.946	0.111	14.872	13.122	0.986	-1.127
ChildVOLat	62	288	21.53	48	0.873	4.245	0.059	11.388	10.458	0.972	-0.838
ChildVOLback	58	545	10.64	31	0.764	3.362	0.208	10.122	10.517	0.967	-0.736
ChildVOLby	33	69	47.83	30	0.883	4.133	0.020	14.851	11.683	0.989	-1.402
ChildVOLdown	91	800	11.38	53	0.889	4.100	0.684	9.810	11.336	0.986	-0.853
ChildVOLfrom	38	127	29.92	32	0.795	3.602	0.049	13.407	11.313	0.908	-0.906
ChildVOLhere	67	324	20.68	51	0.721	4.045	0.046	10.644	9.652	0.984	-0.871
ChildVOLin	186	3647	5.1	68	0.809	3.419	3.945	6.475	9.881	0.984	-0.747
ChildVOLoff	30	119	25.21	25	0.901	3.449	0.056	14.205	12.429	0.921	-0.861
ChildVOLon	132	2602	5.07	44	0.768	3.090	2.413	6.359	9.090	0.977	-0.706
ChildVOLout	105	1562	6.72	45	0.905	3.684	1.051	7.871	10.204	0.963	0.694
ChildVOLover	76	399	19.05	57	0.825	4.396	0.197	11.350	11.665	0.985	-0.915
ChildVOLthere	84	574	14.63	56	0.770	3.815	0.143	9.984	10.641	0.979	-0.842
ChildVOLto	63	682	9.24	33	0.806	3.940	0.498	9.988	10.868	0.937	-0.748
ChildVOLup	141	1619	8.71	77	0.925	4.653	2.105	9.171	11.949	0.965	-0.831
Mean Child VL Class	76.188	845.750	18.441	43.438	0.846	3.819	0.727	10.929	10.966	0.967	-0.784

#5 Semantic Network Tables:

VAC: VOL Construction	**Three Highest Frequency Verbs**	**Three highest Degree Verbs**	**Degree of Frequent Verb Not in Top 3 Degree**	**Total # of Network Nodes**
CDS: V n off n	take (40.3%), get (13.8%), pull (5.1%)	take (37), move (29), get (25)	pull (24)	51
CHI: V n on n	put (56.9%), get (7.9%), have (5.5%)	take (62), get (53), change (52)	put (32), have (34)	96
CDS: V n from n	get (25.8%), take (10.6%), buy (6.6%)	take (51), move (42), draw (37)	get (32), buy (19)	75
CHI: V n from n	get (40.9%), take (11.8%), drink (3%)	take (29), go (19) get (18)	get (17), drink (3)	36

VAC: Clausal Core	Three Highest Frequency Verbs	Three highest Degree Verbs	Degree of Frequent Verb Not in Top 3 Degree	Total # of Network Nodes
CDS: Transitive	*put* (9.6%), *get* (6.7%), *have* (6.3%)	take (63), make (50), draw (50)	put (36), get (47), have (37)	95
CHI: Transitive	*want* (10.1%), *get* (10.1%), *put* (7.3%)	take (62), draw (57), move (52)	get (49), put (37), want (22)	96
CDS: Intransitive	*go* (32.3%), *let* (11%), *know* (9%),	go (58), work (52), move (50)	tell (12), know (10)	96
CHI: Intransitive	*go* (17.42%), *know* (7.7%), *want* (7.5%)	go (55), run (52), draw (52)	want (25), know (7)	97
CDS: Ditransitive	*give* (42.7%), *get* (16.6%), *tell* (8.6%)	give (13), get (13), draw (13)	tell (5)	17
CHI: Ditransitive	*give* (43.2%), *get* (17%), *make* (8.3%)	give (9), get (8), make (8)	n/a	11

VAC: VL Construction	**Three Highest Frequency Verbs**	**Three Highest Degree Verbs**	**Degree of Frequent Verb Not in Top 3 Degree**	**Total # of Network Nodes**
CDS: V *over* n	come (23.5%), go (18.5%), fall (10.5%)	move (62), go (61), run (56)	come (47), fall (48)	97
CHI: V *through* n	go (41.9%), get (14.1%), come (14.1%)	run (16), go (15), get (15)	come (12)	24
CDS: V *around* n	turn (32.7%), go (21.6%), run (8%)	move (58), turn (47), run (46)	go (42)	63
CHI: V *out* n	get (26.9%), come (22.4%), go (12.3%)	go (58), take (56), come (45)	get (40)	97

Bibliography:

Abbot-Smith, K., Lieven, E. and Tomasello, M. (2004), *Training 2;6-year-olds to produce the transitive construction: the role of frequency, semantic similarity and shared syntactic distribution.* Developmental Science, 7: 48–55. doi: 10.1111/j.1467-7687.2004.00322.x

Casenhiser, D. and Goldberg, A. E. (2005) *Fast mapping of a phrasal form and meaning*", Developmental Science

Chomsky, Noam. (1995). *The Minimalist Program*. Cambridge, Mass.: The MIT Press.

C. Lefebvre & H. Cohen (2005), *Handbook of Categorization in Cognitive Science*. New York: Elsevier Science.

Ellis, N. C. (2002). Frequency effects in language acquisition: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition, 24*, 143-188.

Ellis, N. C., & Ferreira-Junior, F. (2009). Constructions and their acquisition: Islands and the distinctiveness of their occupancy. *Annual Review of Cognitive Linguistics*, 111-139.

Ellis, N. C. & O'Donnell, M. (2011). *Robust Language Acquisition – an Emergent Consequence of Language as a Complex Adaptive System*. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), Proceedings of the 33rd Annual Conference of the Cognitive Science Society (pp. 3512-3517). Austin, TX: Cognitive Science Society.

Francis, G., Hunston, S., & Manning, E. (Eds.). (1996). *Grammar Patterns 1: Verbs. The COBUILD Series*. London: Harper Collins.

Goldberg, Adele E. (1995) *Constructions: a Construction Grammar Approach to Argument Structure*. Chicago: University of Chicago.

Goldberg, A. E. (2003). Constructions: a new theoretical approach to language. *Trends in Cognitive Science*, 7, 219-224.

Goldberg, A. E. (2006). *Constructions at work: The nature of generalization in language*. Oxford: Oxford University Press.

Gries, Stefan Thomas, and Anatol Stefanowitsch. *Corpora in Cognitive Linguistics: Corpus-based Approaches to Syntax and Lexis*. Berlin: Mouton De Gruyter, 2006. Print.

Levin, Beth. *English Verb Classes and Alternations: a Preliminary Investigation*. Chicago: University of Chicago, 1993. Print.

Levin, B. and Rappaport Hovava, M. (1993), *The dative alternation revisisted*. Paper presented at the workshop on verb classes and Alternations, Institute Fur Linguistin, Universitat Stuttgar, Jan. 10-11

MacWhinney, Brian. (2000). The CHILDES project: Tools for analyzing talk. Third Edition. Mahwah, NJ: Lawrence Erlbaum Associates

Miller, G. A. (2009). WordNet - About us. Retrieved January 9, 2011, from http://wordnet.princeton.edu

Murphy, G. L. (2002), *The Big Book of Concepts*. Cambridge, MA: MIT Press

Ninio, Anat. *Syntactic Development, Its Input and Output*. Oxford: Oxford UP, 2011. Print.

Pedersen, T., Patwardhan, S., & Michelizzi, J. (2004).*WordNet::Similarity – Measuring the Relatedness of Concepts.* Proceedings of Fifth Annual Meeting of the North American Chapter of the Association of Computational Linguistics (NAACL 2004).

Saffran, J. R., Aslin, R. Newport, E. (1996), *Statistical learning by* 8-month old infants", *Science* 274: 1926-8

Shanks, D. R. (1995). *The psychology of associative learning*. New York: Cambridge University Press.

Solé, R. V., Murtra, B., Valverde, S., & Steels, L. (2005). Language Networks: their structure, function and evolution. *Trends in Cognitive Sciences, 12*.

Tomasello 2005 Constructing a Language: A Usage-Based Theory of Language Acquisition.

Zipf, G. K. (1949). *Human behaviour and the principle of least effort: An introduction to human ecology*. Cambridge, MA: Addison-Wesley.