2016-03-16

Modeling Children's Organization of Utterances Using Statistical Information from Adult Language Input

Katie Lynn Walker
Brigham Young University - Provo

Modeling Children's Organization of Utterances Using Statistical Information from Adult Language Input

Katie Lynn Walker

A thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Science

Ron W. Channell, Chair
Shawn L. Nissen
Christopher Dromey

Department of Communication Disorders

Brigham Young University

March 2016

ABSTRACT

Modeling Children's Organization of Utterances Using Statistical Information From Adult Language Input

Katie Lynn Walker
Department of Communication Disorders, BYU
Master of Science

Previous computerized models of child language acquisition have sought to determine how children acquire grammatical word categories (GWCs). The current study seeks to determine if statistical structure can be corroborated as a factor in GWC acquisition. Previous studies examining statistical structure have dealt with word order rather than GWC order and only examined an overall success rate. The present study examines how well a computer model of child acquisition of GWCs was able to reorganize scrambled sentences back into the correct GWC order using transitional probabilities extracted from adult language input. Overall, a 50% success rate was obtained, but when broken down by utterance length, utterances up to eight words in length had a success rate much higher than chance. Thus, it is likely that statistical structure informs children's acquisition of GWCs.

Keywords: grammatical word categories, statistical structure, language acquisition

ACKNOWLEDGEMENTS

I would like to thank Dr. Channell for his endless patience, help and guidance throughout the process of writing this thesis. Without him, this project would not have been possible. I would also like to thank my family, especially my husband and my parents, for their continual support.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

DESCRIPTION OF THESIS STRUCTURE

This thesis, *Modeling Children's Organization of Utterances Using Statistical Information from Adult Language Input*, is part of a larger research project, and all or part of the data from this thesis may be published as part of articles listing the thesis author as a co-author. The thesis itself is to be submitted to a peer-reviewed journal in speech-language pathology. An annotated bibliography is presented in Appendix A.

Introduction

By the end of the preschool years, children have acquired at least the major grammatical word categories (GWCs) such as nouns and verbs and can apply them to novel words (Berko, 1958). In regard to this acquisition, two main approaches have developed in the literature that attempt to explain how this phenomenon occurs. The nativist approach argues that children are born with an innate knowledge of syntactic categories. On the other hand, the constructivist approach contends that children are not born with these categories, but rather extract them from the language they hear as they learn language. Both theories assume that children have a sophisticated processing mechanism in order to make sense of the language input they are exposed to. Both of these theories have spawned what are called bootstrapping models.

The term bootstrapping refers to the process by which children use one aspect of a language to decipher another (Karmiloff & Karmiloff-Smith, 2002). A variety of bootstrapping models have been proposed in order to provide more information and insight into how children are able to acquire language. These models could be semantic, prosodic, or syntactic. Semantic bootstrapping refers to how children use word meaning to decipher GWCs, such as noting that most words for objects are nouns, most actions are verbs, and so on. Prosodic bootstrapping refers to how children use the intonation patterns of language to decipher syntactic boundaries, such as noting that the noun form of a word such as *convert* has the accent on the first syllable and the verb form of *convert* has the second syllable accented. Syntactic bootstrapping refers to how children use syntax to derive word meaning by noting the distribution of a word relative to its use before, after, or between other words.

In examining these theories, Pinker (1988) noted that "descriptions of the language acquisition process" were often "vague and metaphorical" (p. 98). In order to address this

problem in the study of children's acquisition of language, several researchers have developed computer models to bring the study of language acquisition from the realm of speculation into a context that relies on more quantifiable models.

One such computerized syntactic bootstrapping approach, proposed by Reddington, Chater, and Finch (1998), examined the language that children were exposed to with a distributional information model. Distributional information refers to the linguistic context in which a word occurs. Their report acknowledged the lack of research in computational approaches to children's language acquisition, particularly into how children acquire syntactic categories. By examining distributional information, the authors looked at the input children receive from their environment to explain the acquisition of GWCs. Taking adult language samples from the CHILDES database (MacWhinney, 2000), Redington et al. assigned target words (the 1,000 most frequent words) and context words (the 150 most frequent words) to their most common syntactic category. These results were scored in terms of accuracy, completeness, and informativeness. The nearer a context word was to a target word, the more informative it was regarding the word's grammatical category. While the authors conceded that distributional information does not explain the acquisition of GWCs in its entirety, they concluded that distributional information gave insight into the process of acquiring syntactic categories.

A similar study performed by Mintz (2003) examined how frequent frames, or two words that frequently occur together with another word in between, gave insight into children's acquisition of GWCs. Using corpora from the CHILDES database, Mintz conducted two experiments. In the first one, he used the 45 most frequent frames of each language sample to categorize words. In this experiment, Mintz found that frequent frames were an effective method for categorizing words. This experiment yielded high accuracy. In the second experiment, Mintz

analyzed the frequency relative to the total number of frames in a sample. This experiment was also done to ensure that the high accuracy obtained in the first experiment was not due to a small number of GWCs. However, the frequent frames, though effective for isolating some GWCs, were able to cover only a small portion of the entire language. Perhaps the child's discovery of frequent frames teaches the child enough about GWCs to seek categories even outside the frequent frames that can be statistically isolated.

Stenquist (2015) added a unique perspective to computerized models of child language GWC acquisition by using an evolutionary model. The computer model used in her study employed an adaptation and selection model to evolve a set of GWCs given exposure to language input. Using adult utterances to train the computer model and children's utterances to test the model, the study found a rapid increase in accuracy in generations 1-1500, and a gradual increase in accuracy until generation 12,000 in each of the children's language samples. Her study found that using an evolutionary model was successful in assigning words to grammatical categories even when starting from a random assignment of words to GWCs. This study also provided a unique contribution by evaluating the learnability of a child's language from his or her own parent or caregiver language input.

With these insights in mind, the current investigation seeks to add new perspective to children's acquisition of GWCs through a probabilistic approach. The purpose of the present study is to propose that like other types of bootstrapping, statistical structure could provide further information to a child learning language. This approach has been suggested as plausible by recent research evidence that has acknowledged the applicability of probabilistic approaches to learning and to language acquisition.

Probabilistic approaches to language stand in contrast to categorical approaches. In recent years, more research has been done and a stronger case has been made for a more probabilistic approach. In a book on probabilistic approaches to linguistics, Manning (2003) points out that "human cognition has a probabilistic nature: we continually have to reason from incomplete and uncertain information about the world, and probabilities give us a well founded tool for doing this" (p. 290). Manning goes on to further differentiate a categorical and probabilistic approach to syntax, pointing out that a purely categorical approach is often too simplified, while a probabilistic approach can help to account for the many complexities of human language.

As computerized models have become more widely used, probabilities have been used to further improve computer models and make them more precise. This has gained more attention in recent years with the increase of computerized recognition of speech, which has suggested that there is much to be gained from further exploration of probabilistic approaches to language in general. This has led in turn to studies aimed to examine whether or not probability plays a role in language acquisition, such as learning about a child's acquisition of GWCs.

Clark, Giorgolo, and Lappin (2013) sought to distinguish "whether linguistic knowledge is probabilistic or categorically rule-based in nature" (p. 2064). In the study, a binary classifier was used to compare sentences from the British National Corpora and their reversed counterparts as well as the original sentence with all of its possible variants by randomly exchanging words in each sentence with another word three positions away. These sentences were then scored as either "well-formed," or "distorted." Their findings revealed that computers were able to separate well-formed vs. randomized sentences, thus suggesting a strong correlation between probabilities and grammatical judgments. The authors noted that while probability cannot account for a

speaker's entire grammatical knowledge, a strong correlation was found, so it is likely that probability plays a role.

Other research has been done that acknowledges the reality of probabilistic language with more sophisticated adult language models. Chang, Lieven, and Tomasello (2006) sought to examine if children's language could be explained using the statistical structure gleaned from a training corpus. Their study used child utterances with a computational model to examine word order in twelve typologically different languages. Chang et al. used an evaluation measure based on sentence production models called the Word Order Prediction Algorithm (WOPA). Starting with an unordered set of words, the program predicts correct word order, which is then compared to the original sentence from the corpora. The study's WOPA measures were successful in using different algorithms to determine word order in different languages, which provided a basis for assuming probabilistically-organized structure in children's utterances. One weakness of this study is that it examined superficial word order rather than GWC order, limiting the generalization to new words.

The current study will extend this line of research by examining the role of probability in GWC order in children's utterances. By using only statistically derived measures to reorganize GWC order in children's language, statistical structure could be corroborated as a factor that might help children to learn GWCs.

Method

Training Corpus

Eight sets of spoken language samples (the Adam, Anne, Aran, David, Naomi, Nina, Peter, and Sarah corpora) were taken from the CHILDES database (MacWhinney, 2000) and each divided into two subcorpora: adult utterances spoken to the child and utterances spoken by

the child. The adult utterances spoken to the child were used to train the computer model for the current study. These corpora were used previously in studies by Cluff (2014) and Berardi (2015). No information is available regarding the socio-economic status of the families of Anne, David, Naomi, and Nina.

Adam (Brown, 1973) Adam came from a family described as middle-class and well educated. A total of 55 files of Adam's spontaneous speech were recorded from age 2;3 (years;months) to age 4;10. A total of 19,301 adult utterances from this set of language samples were used as training utterances for the current study.

Anne (Sawyer, 1997) Anne was recorded at age 3;5 in her preschool classroom during a study that examined how unstructured play contributes to conversational skills, social skills and creativity development. A total of 25,551 adult utterances from this set of language samples were used as training utterances for the current study.

Aran (Theakston, Lieven, Pine, & Rowland, 2001) Aran was the oldest child from a middle-class family. He was recorded as part of a study of children's acquisition of verb-argument structure in his home twice every three weeks over a one-year period. A total of 20,192 adult utterances from this set of language samples were used as training utterances.

David (Henry, 1995) David was the oldest child in his family who was recorded from age 2;0 to 4;2. His language samples were recorded as part of a study examining how children acquiring English in Belfast, Northern Ireland use variable subject-verb agreement. A total of 9,933 adult utterances from this set of language samples were used as training utterances in the current study.

Naomi (Sachs, 1983) Naomi was recorded by her mother as part of a longitudinal study from age 1;1 to 5;1. A total of 12,034 adult utterances from this set of language samples were used as training utterances in the current study.

Nina (Suppes, 1974) Nina was recorded as part of a study of semantics in children's speech from age 1;11 to 3;3. A total of 35,381 adult utterances from this set of language samples were used as training utterances for the current study.

Peter (Bloom, Hood, & Lightbown, 1974; Bloom, Lightbown & Hood, 1975) Peter was the oldest child from an upper-middle-class family. He was recorded from age 1;9 to 3;2 as a part of two studies that examined the role of imitation in language development and structure and variation in child language. A total of 48,205 adult utterances from this set of language samples were used as training utterances in the current study.

Sarah (Brown, 1973) Sarah came from a working-class family. A total of 139 files of Sarah's spontaneous speech were recorded from age 2;3 to age 5;1. A total of 48,205 adult utterances from this set of language samples were used as training utterances in the current study.

Test Corpus

The last 500 of each child's utterances, excluding single-word utterances, were used as the test corpora for the current study. For David's corpus, however, only 427 utterances were available for use.

Computer Model

The simulation software entitled tt_m3.jl (Channell, 2015) was used in the current study. Before the program's use, a dictionary is compiled which contains the words in the training and test corpora and the most common grammatical category tag for each word. The program starts by reading the training corpus, grammatically tagging each word using the dictionary, and

extracting the tag transition frequencies of lengths one, two, and three from each utterance in the tagged training corpus. The test corpus is then read and each utterance within it is grammatically tagged using the dictionary, and this resulting tag sequence is saved as the original string. These tags are then scrambled to create a set of tags to be reordered. To reorder the tags, the transition frequencies are used to make a list of the most probable tag sequences starting with each tag. For utterances that are two, three, or four words long, the highest probability sequence option on this list is taken and compared to the original tag string to judge correctness. For utterances that are N = 5 words or longer, if one of the N possibilities matches the original sequence, it is judged as correct. This slightly wider tolerance is allowed because of the very low chance likelihood of longer utterances. For example, an eight word utterance has 8! (eight factorial) = 40,320 possible tag orderings, and the program reconstructed eight (one starting with each tag) tag orderings; if one of those eight orderings matched the original tag sequence, it was counted as correct. Finally, the program writes the results organized by utterance length to a file.

The following steps outline the computer model:

1. A corpus of adult/child conversation is downloaded from the CHILDES database.
2. Unneeded information is removed, just leaving the adult and child utterances.
3. The adult utterances are placed in a file.
4. The last 500 child utterances are placed in a different file.
5. The words in the adult utterances are given their most likely grammatical category tag, based on the dictionary used in Channell and Johnson (1999).
6. The transition probabilities are noted for pairs and trios of these grammatical category tags.
7. The words in the 500 child utterances are given their most frequent grammatical

category tag, based on the same dictionary used for the adult utterances.

8. Each child utterance is disassembled and then reassembled using the adult sample transition probabilities.

9. The number of exact matches of the reassembled utterances to the original child utterances is recorded.

Procedure

All eight children's language corpora, originally from the CHILDES database as described above, were previous used in studies done by Cluff (2014) and Berardi (2015) in which punctuation was removed. This corpus was further modified for the current study by dividing the utterances into c-units (communication units), which are each of the independent clauses in a sentence along with any dependent clauses. After the utterances were divided into c-units, a corpus of adult utterances was run through the algorithm to train the computer model. After the computer model was trained on the adult utterances, each child's language sample was individually run through the computer program, and the number of child utterances that could be correctly reorganized by the GWC transition probabilities extracted from the adult utterances was measured. The study's measure of interest was, out of each sample's 500 utterances, how many were correctly reassembled using transition probabilities (the dependent variable) as a function of utterance length (an independent variable).

Results

After each child's language sample was run through the computer program, the number of child utterances that could be correctly reorganized using transition probabilities from the adult training corpus was calculated as a percentage for each utterance length. The average percentages are shown in Table 1. These percentages were highest for shorter utterances and

decreased for longer utterances, as would be expected. However, up to utterance lengths of eight words or fewer, the observed percentages were consistently higher than the levels expected by chance. A one-sample *t*-test was performed to compare the differences between mean percent correct and the expected percent correct, and these differences were found to be statistically significant.

Table 1

Utterance Reorganization Performance and Comparison to Expected Values by Utterance Length, Across All Children

Length	M	*SD*	Expected	*t*
2	79.15	5.43	50.00	15.17**
3	58.79	8.79	16.67	13.56**
4	40.93	7.06	4.16	14.73**
5	38.39	4.88	0.83	21.77**
6	28.07	10.34	0.14	7.64**
7	13.90	13.71	0.02	2.86*
8	7.74	6.92	0.00	3.16*
9	1.39	3.93	0.00	1.00

Note. The actual expected level for utterance length 8 is 0.00248 and for length 9 is 0.0002758.
** $p < .01$; * $p < .05$

The reorganization using transition probabilities worked better for the samples of some adults and children than others. Tables 2 through 8 display each child's results. It can be seen in these tables that the overall levels of correct utterance reorganization ranged from 33.80% to 48.80%, that utterances five words long were typically 35% higher than would be expected, and that utterances longer than nine words were not correctly reorganized. The percent accuracy for

an utterance length of 2 ranged from 72.31 to 88.10, for a length of 3 ranged from 42.27 to 67.95, for a length of 4 ranged from 26.97 to 48.62, and accuracy for an length of 5 ranged from 34.29 to 49.41. The accuracy for a length of 6 ranged from 11.29 to 42.31; accuracy for a length of 7 ranged from 0.00 to 43.24. The percent accuracy for a length of 8 ranged from 0.00 to 23.53 and for a length of 9 ranged from 0.00 to 11.11. Some corpora did not contain any utterances of length 7, 8 or 9, which accounts for the percentages of 0.00 in those corpora.

Table 2

Utterance Length, Percent Correct, and Expected Percent Correct for Adam's Corpus

Length	N Correctly Reorganized	N Utterances	Observed Percent	Expected Percent
2	37	42	88.10	50.00
3	53	78	67.95	16.67
4	53	109	48.62	4.16
5	30	86	34.88	0.83
6	16	67	23.88	0.14
7	4	47	8.51	0.02
8	2	36	5.56	0.00
9	0	15	0.00	0.00
Total	195	500	39.00	7.87

Note. A total of 20 utterances 10 words or longer were not correctly reorganized. The expected level for utterance length 8 is 0.00248 and for length 9 is 0.000276.

Table 3

Utterance Length, Percent Correct, and Expected Percent Correct for Anne's Corpus

Length	N Correctly Reorganized	N Utterances	Observed Percent	Expected Percent
2	75	89	84.27	50.00
3	66	125	52.80	16.67
4	48	106	45.28	4.16
5	29	82	35.37	0.83
6	14	57	24.56	0.14
7	2	19	10.53	0.02
8	1	13	7.69	0.00
9	0	6	0.00	0.00
Total	235	500	47.00	14.10

Note. A total of 3 utterances 10 words or longer were not correctly reorganized. The expected level for utterance length 8 is 0.00248 and for length 9 is 0.000276.

Table 4

Utterance Length, Percent Correct, and Expected Percent Correct for Aran's Corpus

Length	N Correctly Reorganized	N Utterances	Observed Percent	Expected Percent
2	47	65	72.31	50.00
3	41	97	42.27	16.67
4	44	123	35.77	4.16
5	36	105	34.29	0.83
6	7	62	11.29	0.14
7	0	32	0.00	0.02
8	0	16	0.00	0.00
9	0	0	0.00	0.00
Total	175	500	35.00	10.95

Note. The expected level for utterance length 8 is 0.00248 and for length 9 is 0.000276.

Table 5

Utterance Length, Percent Correct, and Expected Percent Correct for David's Corpus

Length	N Correctly Reorganized	N Utterances	Observed Percent	Expected Percent
2	41	54	75.93	50.00
3	29	54	53.70	16.67
4	38	99	38.38	4.16
5	27	73	36.99	0.83
6	22	52	42.31	0.14
7	6	42	14.29	0.02
8	2	26	7.69	0.00
9	1	9	11.11	0.00
Total	166	427	38.88	9.56

Note. A total of 18 utterances 10 words or longer were not correctly reorganized. The expected level for utterance length 8 is 0.00248 and for length 9 is 0.000276.

Table 6

Utterance Length, Percent Correct, and Expected Percent Correct for Naomi's Corpus

Length	N Correctly Reorganized	N Utterances	Observed Percent	Expected Percent
2	29	38	76.32	50.00
3	41	70	58.57	16.67
4	24	89	26.97	4.16
5	40	108	37.04	0.83
6	26	71	36.62	0.14
7	8	51	15.69	0.02
8	1	23	4.35	0.00
9	0	22	0.00	0.00
Total	169	500	33.80	7.08

Note. A total of 28 utterances 10 words or longer were not correctly reorganized. The expected level for utterance length 8 is 0.00248 and for length 9 is 0.000276.

Table 7

Utterance Length, Percent Correct, and Expected Percent Correct for Nina's Corpus

Length	N Correctly Reorganized	N Utterances	Observed Percent	Expected Percent
2	51	64	79.69	50.00
3	58	88	65.91	16.67
4	54	116	46.55	4.16
5	42	85	49.41	0.83
6	19	66	28.79	0.14
7	16	37	43.24	0.02
8	4	17	23.53	0.00
9	0	8	0.00	0.00
Total	244	500	48.80	10.46

Note. A total of 19 utterances 10 words or longer were not correctly reorganized. The expected level for utterance length 8 is 0.00248 and for length 9 is 0.000276.

Table 8

Utterance Length, Percent Correct, and Expected Percent Correct for Peter's Corpus

Length	N Correctly Reorganized	N Utterances	Observed Percent	Expected Percent
2	56	68	82.35	50.00
3	53	84	63.10	16.67
4	52	125	41.60	4.16
5	36	92	39.13	0.83
6	24	64	37.50	0.14
7	0	34	0.00	0.02
8	1	12	8.33	0.00
9	0	7	0.00	0.00
Total	222	500	44.40	10.81

Note. A total of 14 utterances 10 words or longer were not correctly reorganized. The expected level for utterance length 8 is 0.00248 and for length 9 is 0.000276.

Table 9

Utterance Length, Percent Correct, and Expected Percent Correct for Sarah's Corpus

Length	N Correctly Reorganized	N Utterances	Observed Percent	Expected Percent
2	46	62	74.19	50.00
3	68	103	66.02	16.67
4	50	113	44.25	4.16
5	36	90	40.00	0.83
6	11	56	19.64	0.14
7	7	37	18.92	0.02
8	1	21	4.76	0.00
9	0	9	0.00	0.00
Total	195	500	43.80	10.74

Note. A total of 9 utterances 10 words or longer were not correctly reorganized. The expected level for utterance length 8 is 0.00248 and for length 9 is 0.000276.

In an attempt to understand why the reorganization using transitional probabilities performed better for some children's language samples than for others, the following variables were examined: (a) total number of adult utterances in the training model, (b) total number of adult words, (c) adult MLU (mean length of utterance, in words), (d) child MLU, and (e) total number of child words. It may be recalled that each child's corpus was 500 utterances long except for David's, for which only 427 utterances had been available. In Table 10, each of these variables is detailed for each adult and each child corpus.

Table 10

Number of Utterances and Words and MLU Levels for Adult and Child Samples

Child	Total number of adult utterances	Total number of adult words	Adult MLU	Total number of child words	Child MLU
Adam	17151	90944	5.30	2570	5.14
Anne	18469	94784	5.13	2051	4.10
Aran	18535	113946	6.15	2159	4.32
David	9146	56763	6.21	2125	4.98
Naomi	10057	55107	5.48	2663	5.33
Nina	33422	193127	5.78	2354	4.71
Peter	16957	93538	5.52	2282	4.56
Sarah	36444	181903	4.99	2279	4.56

In examining each of these variables, a correlational coefficient (r) was calculated to determine whether a significant correlation existed between the independent variable (total number of adult utterances, total number of adult words, etc.) and the dependent variable of overall percentage correct of reorganizing utterances. Table 11 displays the r value for each of the independent variables listed above in Table 10. However, with only 6 degrees of freedom, none of these correlations reached statistical significance, as a correlation of .707 is needed to reach the $p < .05$ level and a correlation of .834 would be needed to reach the $p < .01$ level.

Table 11

Correlations Between Linguistic Characteristics of the Language Samples and Utterance Reorganization Performance

Linguistic variable	*r*
Total number of adult utterances	0.63
Total number of adult words	0.61
Adult MLU	-0.38
Child MLU	-0.51
Total number of child words	-0.39

Note. d.f. = 6

Discussion

The purpose of the current study was to examine how well a computer program that modeled a child's acquisition of GWCs was able to reorganize scrambled sentences back into the correct GWC order using transitional probabilities extracted from adult language input. In so doing, the current investigation sought to determine if statistical structure could be corroborated as a factor that might help children to learn GWCs and thus provide greater insight into the process of GWC acquisition. In general, the computer model was able to reorganize utterances of eight words or fewer at a much higher percentage than would be expected by chance. This suggests that statistical structure may play a role in the child's acquisition of the ability to formulate utterances.

The findings of the present study are comparable to the results of similar studies. Chang et al. (2006) used an algorithm to examine syntax acquisition using word order in twelve typologically different languages. Starting with an unordered set of the words from an utterance, the program attempted to predict the correct word order using probabilities. These newly formed

word sequences were then compared to the actual sentences from the corpora. The study by Chang et al. demonstrated that their utterance reordering measure was successful in quantifying the use of five different reorganization algorithms in twelve different languages, providing a computational model that could be applied to different languages. The present study used the utterance reorganization task proposed by Chang et al. but provided additional insight by examining GWC order rather than word order. The use of GWCs instead of the use of specific words offers a means whereby children can generalize utterance reorganization to new words. Wintner (2010) examined the study by Chang et al. and pointed out that the utterances used in that study were all very short, averaging less than three and one half words per utterance and that the study of Chang et al. also only examined the overall success rate. The current paper added a new feature to this line of research by not only examining the overall success of the computer model, but also by breaking down success rates by utterance length. Of course, the use of grammatical categories limits the findings of the present paper just to English, rather than the multiple languages examined in the Chang et al. study.

A similar and more recent study by Clark, Giorgolo, and Lappin (2013) examined a statistical model of grammaticality with adult language. They presented scrambled and well-formed adult utterances to a computer model which then classified them as either “well-formed” or “distorted” with a very high success rate (p. 2064). They concluded that while linguistic knowledge cannot be simply reduced to probability, there is still a strong correlation between the two. The present study, rather than simply identifying an utterance as acceptable or not, attempted to reorganize the utterance into a satisfactory form. The present study also extended this line of research by applying it to child language acquisition.

Another similar study for which the present results have implications is that of Mintz (2003). Mintz had examined how frequent frames, or two words that frequently occur together with another word in between, gave insight into children's acquisition of GWCs. Using corpora from the CHILDES database, Mintz conducted experiments which showed that the 45 most frequent frames of each language sample could be used to categorize words with high accuracy. However, the frequent frames, though effective for isolating some GWCs, were able to cover only a small portion of the entire language. Perhaps as Mintz suggested, the child's discovery of frequent frames teaches the child enough about GWCs to seek categories even outside the frequent frames that can be statistically isolated. The findings of the present study suggest that three-word sequences, as used by Mintz, contain enough probability information to allow much better than chance reorganization of utterances up to nine words in length.

The present study has several limitations. With only eight corpora of adult utterances to train the model and eight child corpora to test the model, the study is constrained in how broadly the results can be applied. Future research should examine not only a greater number of corpora, but also a broader range of samples that reflect a greater range of people. This could include examining child language samples from various ethnicities, SES (socioeconomic status) groups, etc. Future research should also examine applicability to other languages, as the current investigation only examined corpora from English speakers.

Another limitation is that this study used only one set of grammatical tags. The dictionary of the GramCats program (Channell & Johnson, 1999) was used as the basis for tagging the words in this study; this dictionary contains 85 different grammatical tags. The computer program had to determine transitional probabilities for 85 possible GWCs. In comparison, the LARSP analysis procedure has 22 word-level grammatical tags (Crystal, Garman, & Fletcher,

1989). In Mintz's 2003 study, a set of 11 grammatical tags and a set of 14 grammatical tags were both used, and the results using both different tag sets were compared. Because only one set of grammatical tags was used in the present study, the effects of a certain number of tags on the transitional probabilities is unknown. Broader tagging may have yielded different results in the overall success rate. Thus, future research should examine the effects of using different-sized sets of grammatical tags.

Future research should also examine factors that cause some child language samples to score higher than others, in that utterances were more likely to be precisely reorganized. In the current study, the factors of (a) total number of adult utterances in the training model, (b) total number of adult words, (c) adult MLU (mean length of utterance, in words), (d) child MLU, and (e) total number of child words were examined as potential contributors to performance, but none were determined to have statistically significant relationships. It is thus still uncertain as to why some children's samples performed better than others. This could be due to the small sample size, but the various performance rates could also be due to other factors not examined in this study. Future research should also examine factors such as the diversity of GWCs to which the child is exposed, the amount of repetitious content to which the child is exposed, gender differences, or other such factors.

Nevertheless, this study contributed a new perspective to the ongoing efforts to determine which factors inform child language learning. This is the first study to examine reorganization of utterances by GWC and to divide the discussion of results by utterance length. When compared with levels expected by chance, the computer model was able to reorganize utterances using transitional probabilities at a much higher percentage. This suggests that the statistical structure

of language the child is exposed to could serve to inform their acquisition of both GWCs and the sequencing of those GWCs.

References

Berardi, E. (2015). *A model of children's acquisition of grammatical word categories from adult input using an adaptation and selection algorithm.* (unpublished master's thesis). Brigham Young University, Provo, UT.

Berko, J. (1958). The child's learning of English morphology. *Word*, *14*, 150-177.

Bloom, L., Hood, L., & Lightbown, P. (1974). Imitation in language development: If, when and why. *Cognitive Psychology, 6*, 380-420.

Bloom, L., Lightbown, P., & Hood, L. (1975). Structure and variation in child language. *Monographs of the Society for Research in Child Development, 40* (Serial No. 160).

Brown, R. (1973). *A first language: The early stages*. Cambridge, MA: Harvard University Press. doi:10.4159/Harvard.9780674732469

Chang, F., Lieven, E., & Tomasello, M. (2006). Using child utterances to evaluate syntax acquisition algorithms. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society.* Vancouver, Canada.

Channell, R. W. (2015). tt_m3.jl [Computer software]. Provo, UT: Brigham Young University.

Channell, R. W., & Johnson, B. W. (1999). Automated grammatical tagging of child language samples. *Journal of Speech, Language, and Hearing Research, 42,* 727-734. doi:10.1044/jslhr.4203.727

Clark, A., Giorgolo, G., & Lappin, S. (2013). Towards a statistical model of grammaticality. In *Proceedings of the 35th Annual Conference of the Cognitive Science Society 2013.* (pp. 2064-2069). Berlin, Germany: Cognitive Science Society.

Cluff, S. Z. (2014). *A model of grammatical category acquisition using adaptation and selection.* (master's thesis). Brigham Young University, Provo, UT. Retrieved from http://scholarsarchive.byu.edu/etd/4086

Crystal, D., Garman, M., & Fletcher P. (1989). *The grammatical analysis of language disability: A procedure for assessment and remediation* (2nd ed.). London, England: Cole and Whurr.

Henry, A. (1995) *Belfast English and Standard English: Dialect variation and parameter setting*. New York, NY: Oxford University Press.

Karmiloff, K., & Karmiloff-Smith, A. (2002). *Pathways to language: From fetus to adolescent*. Cambridge, MA: Harvard University Press.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk (3rd ed.)*. Mahwah, NJ: Lawrence Erlbaum Associates.

Manning, C. D. (2003). Probabilistic syntax. In S. Jannedy, J. Hay, & R. Bod (Eds.) *Probabilistic linguistics* (pp. 289-341). Cambridge, MA: Bradford.

Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition, 90*, 91-117. doi:10.1016/S0010-2077(03)00140-9

Pinker, S. (1988). Learnability theory and the acquisition of a first language. In F. Kessel (Ed.), *The development of language and of language researchers: Papers presented to Roger Brown* (pp. 97-118). Hillsdale, NJ: Erlbaum.

Redington, M., Chater, N., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science, 22*, 425-469. doi: 10.1207/s15516709cog2204_2

Sachs, J. (1983). Talking about the there and then: The emergence of displaced reference in parent-child discourse. In K. E. Nelson (Ed.), *Children's language* (Vol. 4, pp. 1-28) Hillsdale, NJ: Lawrence Erlbaum Associates.

Sawyer, K. (1997). *Pretend play as improvisation*. Mahwah, NJ: Lawrence Erlbaum Associates.

Stenquist, N. A. (2015). *Modeling children's acquisition of grammatical word categories from adult input using an adaptation and selection algorithm.* (unpublished master's thesis). Brigham Young University, Provo, UT.

Suppes, P. (1974). The semantics of children's language. *American Psychologist, 29*, 103-114.

Theakston, A. L., Lieven, E. V. M., Pine, J. M., & Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *Journal of Child Language, 28*, 127-152.

Wintner, S. (2010). Computational models of language acquisition. In A. Gelbukh (Ed.), *Computational linguistics and intelligent text processing* (pp. 86-99). Berlin, Germany: Springer. doi: 10.1007/978-3-642-12116-6_8

Appendix A: Annotated Bibliography

Arciuli, J., & Torkildsen, J. V. (2012). Advancing our understanding of the link between statistical learning and language acquisition: The need for longitudinal data. Front. Psychol. 3:324. doi: 10.3389/fpsyg.2012.00324

This paper reviews previous research suggesting a link between statistical learning and language acquisition. The authors introduce the debate that has existed in previous research between innateness and learning when it comes to language acquisition. Statistical learning is implicit learning that can be assessed through a sequential learning paradigm. Some characteristics of statistical learning outlined in this paper include the ability to learn non-adjacent patterns, not decaying rapidly and the ability to generalize. Interest in the area of linking statistical learning to language acquisition is increasing. Several studies are outlined in the paper that help link these two. Using various imaging instruments, recent research has found that statistical learning uses the same portion of the brain as language. Taking this research one step further, a study by de Vries et al. (2010) determined a causal relationship between using Broca's area of the brain and learning artificial grammars (a form of statistical learning). Another study examined patients with agrammatic aphasia and their ability to complete a statistical learning task of determining the grammaticality of an artificial grammar they were exposed to which was comprised of non-linguistic symbols. Compared with a control group, the patients with agrammatic aphasia performed poorly. These results suggested that the language impairment in aphasia is related to an impairment in statistical learning. Previous research is outlined that links statistical learning with proficient spoken language. These studies find that adults with proficient language are able to perform well on statistical learning tasks independent of other cognitive factors. Adults with language impairment on the other hand perform poorly on these tasks. While these studies help to gain insight into the relationship between statistical learning and language, they did not examine children, which my thesis examines. Research has also established a relationship between statistical learning and written language. Without explicit instruction, children and adults are able to map the pronunciation of novel words using statistical learning, with higher sensitivity with increased exposure to written language (the child's age). In conclusion, the authors reestablish that previous research has shown a link between statistical

learning and language acquisition—that is, humans are equipped with a mechanism that can decipher statistical regularities. The authors also point out the need for future research in this area—a need for longitudinal studies, and studies that examine statistical learning impairment in populations such as autism, SLI and dyslexia. Consistent with this call for further research in this area, my thesis examines children's statistical learning of grammatical word categories.

Berko, J. (1958). The child's learning of English morphology. *Word, 14*, 150-177.

The author outlines the purpose of the study—to determine if children have a knowledge of morphological rules and can generalize English morphology. This is tested by using nonsense words to assess whether children can apply the rules of English morphology. The author points out that previous research has concluded that children can apply English morphology to real words, but this could be due to rote memorization. By using made-up words, this study tests to see if children possess morphological rules, rather than just memorizing the language they hear. The author created 27 nonsense words with corresponding picture cards. Actual words were included as well. Subjects tested included 12 adults as well as 19 preschool-aged children and 61 elementary-school students from ages 5-7. The article then outlines the 27 different cards that tested morphological knowledge. One such example included "this is a wug. Now there is another one. There are two of them. There are two ___." In the discussion section, the author outlined findings in the following morphological areas: formation of the plural, verb inflections, formation of the possessive, adjectival inflection, derivation and compounding, and analysis of compound words. The author refers back to the question brought up in the introduction—do children possess morphological rules? Berko pointed out that if linguistic knowledge consisted of simply memorized words, then children might refuse to answer the questions asked of them because they had never heard the words before. The author reveals that the children answered the questions, were not always correct as far as English morphology is concerned, yet presented consistent and methodical responses, suggesting that they are operating under a knowledge of English morphology. The study determined no significant differences between genders. It found that first graders performed significantly better on slightly less that half of the tested morphemes. The author suggested that the older children were able to perfect morphological knowledge they already had when they were younger. This study affirms that children have acquired grammatical

categories by the end of the preschool years. My thesis study seeks to use a probabilistic approach to provide a possible picture for how this acquisition occurs in children.

Chang, F., Lieven, E., & Tomasello, M. (2006). Using child utterances to evaluate syntax acquisition algorithms. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society.* Vancouver, Canada.

The authors provide background into the growing interest in computer approaches to syntactic acquisition. This study aims to use child utterances with a computational model to examine word order. In this study they use an evaluation measure based on sentence production models called WOPA. Starting with an unordered set of words, the program predicts correct word order. The newly formed sequence is then compared to the actual sentence from the corpora. This study uses children's utterances from twelve typologically different languages, which may give insight into whether or not this computational model can be universally applied. Other studies have typically only examined English. The next section examines five different categorization algorithms used in predicting word order in the present corpora. These algorithms include lexstat learner, prevword learner, freqframe learner, token/type learner and type/token learner. In using these five different algorithms, the goal is to see which is best able to learn implicit constraints in the utterances found in the corpora. In evaluating these different learners, they found that when the adjacency learner (a two-way Markov model that looks at one word or two words right in front of the target word) and the prominence learner (which gathers the most important information from the front of the sentence) worked best when combined together. The two combined learners were more accurate in predicting word order. This study examined word order rather than syntactic categories of children's language using a computational system. The study demonstrated that WOPA measures were successful in using five different categorization algorithms in twelve different languages, providing a universal applied computational model. This study by Chang, Lieven and Tomasello examined word order of children's language using a computational model. My thesis hopes to extend this research by using a probabilistic computational model to examine order of grammatical categories in children's utterances.

Channell, R. W., & Johnson, B. W. (1999). Automated grammatical tagging of child language samples. *Journal of Speech, Language, and Hearing Research, 42,* 727-734. doi:10.1044/jslhr.4203.727

This study examines a probabilistic approach to tagging child language samples by grammatical word category (GWC). Previous research showed success with adult language samples, having a high accuracy rate for automated tagging when compared to manual tagging. Child language samples, however, present with a unique set of challenges to automated tagging due to speech that has not fully matured. The authors use two forms of probabilistic information (relative tag likelihood and tag transition likelihood) to confidently yield the most accurate results possible. The study compares the accuracy of automated and manual tagging of child language samples. The automated grammatical tagging software used in this study was called GramCats. The automated tagging yielded an overall success rate of 95.1%. This success rate was comparable to those found in previous research of adult language samples. My thesis also uses a probabilistic approach to child language samples, and it uses the grammatical tags from the GramCats software.

Clark, A., Giorgolo, G., & Lappin, S. (2013). Towards a statistical model of grammaticality. In *Proceedings of the 35th Annual Conference of the Cognitive Science Society 2013.* (pp. 2064-2069). Berlin, Germany: Cognitive Science Society.

The authors introduce the divide between two schools of thought: "whether linguistic knowledge is probabilistic or categorically rule-based in nature," (p. 2064). They propose their theory that while linguistic knowledge cannot be simply reduced to probability, there still may be a strong correlation between the two. In examining their theory, they compare sentences from the British National Corpora and their reverse counterparts as well as the original sentences with all of its possible variants by randomly exchanging words in each sentence with another word three positions away. These sentences are then scored using binary classifiers as either "well-formed," or "distorted" (p. 2064). The authors provide some background information for the context of the study by examining the potential relationship between probability and grammaticality and present logarithms that compute the probability of a sentence. The authors concede that

grammaticality cannot be simply reduced to probability, but offer that a relationship may still exist between the two. They cite other studies that indicate that probability affects all domains of cognition, so they now seek to examine the domain of grammaticality. The authors present their strategies and equations used to determine the grammaticality of the original and distorted sentences. After performing their experiments, they found that the percentage of correctly distinguishing between the original and reversed sentences was quite high (98.9%). The percentage of distinguishing between original sentences and ones where words were exchanged with other words in the sentence (permuted sentences), was lower (77.3%). The authors analyze the cases where the binary classifiers were unsuccessful from distinguishing the original from permuted and reversed sentences. They found cases of false ungrammatical sentences, sentences where the original and permuted sentences were identical (because identical words were swapped), as well as many sentences that were "semantically odd, but otherwise well-formed sentences," (p. 2067). The authors reaffirm their original thought that while probability cannot account for a speaker's entire grammatical knowledge, there is a strong correlation between "the probability distribution over the sentences of a language and a speaker's grammaticality judgments," (p. 2068). This topic is being further researched with more sophisticated language models, and evaluating these models against native speakers' acceptability judgments. The approach taken in my thesis will be similar to the approach taken in this article. This approach to grammatical knowledge is being currently researched with more sophisticated language models, but has not yet been applied to children's utterances, which is what I aim to do in my thesis. Using a similar approach in my thesis, I hope to determine the extent to which children's utterances can be probabilistically described.

Conwell, E. & Morgan, J. (2007). Resolving grammatical category ambiguity in acquisition. In H. Caunt-Nulton, S. Kulatilake and I. Woo (Eds.), *Proceedings of the 31st Annual Boston University Conference on Language Development*. Somerville, MA: Cascadilla Press.

This study examines words that can be used in two different grammatical categories (e.g., *I like fish, I can fish*) and how children are able to make sense of words with grammatical ambiguity. Grammatically ambiguous words serve as one of the major critiques of distribution-based models of language acquisition. Such research has examined words that are categorically

unambiguous, thus research is lacking in this area. The authors seek to answer the questions of children's experience with these words, if infants can distinguish acoustical differences in pronunciation of these words used as a noun vs. as a verb and also if children are able to produce these words in different categories. To answer the first question of children's experience with grammatically ambiguous words, the authors examined a corpus of parental speech and calculated how frequently each word was used, separating into high, medium and low frequency groups. Within each frequency group, potentially grammatically ambiguous words were marked with the syntactic category used. With each of these words, the number of times they were used in each syntactic category was calculated. The proportion of words used in both categories was then calculated against total number of potentially ambiguous words. Based on their results, they found that while cross-syntactic category use of words was not as frequent as it could be, children are still exposed to words used across syntactic categories. They conclude that there may be cues beyond distribution that help children distinguish learn these words. To answer their second question of if these cues for grammatically ambiguous words are available to infants, the authors habituated infants to either all nouns or all verbs, then exposed the infants to novel words of the same category followed by novel words from the category whey were not exposed to. The results found that infants looked longer when exposed to the words of the different category than the same, suggesting that infants can distinguish acoustical differences in the pronunciation of grammatically ambiguous words. Using the same corpora from experiment one, the authors examined children's production of potentially ambiguous words. They found that the children's use of ambicategorical words was significantly correlated to their mothers' use. They conclude that children are able to map these words in two distinct grammatical categories rather than a single form with two uses, and this behavior is influenced by the statistics of their language input. Overall, the study found that grammatically ambiguous words do not create a hindrance to distribution-based models of children's acquisition of syntactic categories. This study provided strong evidence for children's language being influenced by the statistics of their language input, which influence the current paper seeks to examine.

Höhle, B. (2009). Bootstrapping mechanisms in first language acquisition. *Linguistics*, 47, 359–382.

In terms of language acquisition, bootstrapping refers to how the child has an inherent mechanism to start the language acquisition process. In examining bootstrapping, researchers look at how speech input and language acquisition interact. This paper examines previous research that shows that children are equipped with bootstrapping mechanisms. The author raises some questions to be answered given the bootstrapping theory: how is the child able to make sense of the specific language they are exposed to and how is the child able to analyze language input without already having some knowledge of the language? With this in mind, bootstrapping methods have the role of using structural properties of language input to act as parameters for further language learning. The author outlines various kinds of bootstrapping methods including distributional bootstrapping, semantic bootstrapping, syntactic bootstrapping, typological bootstrapping and prosodic bootstrapping. Distributional bootstrapping refers to co-occurring features of a language (phonemes, morphemes, syllables, words, etc.) that provide information for syntactic categories. Semantic bootstrapping refers to how children use word meaning to decipher grammatical word categories, such as noting that most words for objects are nouns, most actions are verbs, and so on. Syntactic bootstrapping refers to how children use syntax to derive word meaning. Typological bootstrapping refers to the intersect between different linguistic elements such as semantics and syntax to draw inferences about the meaning of new members of different grammatical word categories. Prosodic bootstrapping refers to how children use the intonation patterns of language to decipher syntactic boundaries. The author then goes on to further detail prosodic bootstrapping. Bootstrapping models have some very clear strengths. These include the ability to make certain predictions about the variety of input the child uses to acquire language. These models provide a natural explanation for a complex process and account for interactions between various aspects of language (semantics, syntax, prosody, etc.) and how they help to acquire another aspect of language. These models also provide insight into children with language impairment, as research has concluded that children with language impairment have impaired bootstrapping ability. Some weaknesses exist too, however. Bootstrapping methods are still unable to answer the question of initially "penetrating"

the language system without any prior linguistic knowledge as an infant. The author concludes to paper by outlining further research to be done in this field. Among her suggestions include examining other cues children may use to complete bootstrapping. In line with this proposition, my thesis seeks to examine how children use statistical structure of language input to complete bootstrapping.

Jurafsky, D. (2003). Probabilistic modeling in psycholinguistics: Linguistic comprehension and production. In S. Jannedy, J. Hay, & R. Bod (Eds.) *Probabilistic linguistics* (pp. 39-95). Cambridge, MA: Bradford.

The author provides some background for research that has been done concerning probability theory, cognition and language. He outlines some of the roles that probability has been claimed to play in language comprehension, production and learning. Probabilistic modeling has been used to describe phonology, morphology, lexical processing, and syntax. This chapter focuses on lexical and syntactical processing. The author summarizes prior research in the field of probabilistic grammar. Much research has been done concerning frequency of lexicon and syntax. By examining the evidence, the author showed that frequency plays a key role in lexicon and syntax both expressively and receptively. In the next section, the author outlines probabilistic architectures for modeling the frequency effects that he discussed in the previous section. These models include constraint-based models, which "focus on the interactions of a large number of probabilistic constraints to compute parallel competing interpretations," the competition model, which "map[s] from the 'formal' level (surface forms, syntactic constructions, prosodic forms, etc.) to the 'functional' level (meanings, intentions)," rational models, which "claim that human cognitive processing makes optimal use of limited resources to solve cognitive problems." The author also outlines more sophisticated probabilistic models, which include Markov models, stochastic context-free grammars, and Bayesian belief networks. The author outlines potential downfalls to probabilistic models as well as answers questions to potential confusions about probabilistic models. The author of this chapter provides the rationale for using a probabilistic model in examining linguistics. As my thesis seeks to use a probabilistic model to describe children's language, the author's rationale warrants further study of this topic.

Kübler, S. (1998). Learning a lexicalized grammar for German. In D. M. W. Powers (ed.) NeMLaP3/CoNLL98: *New methods in language processing and computational natural language learning*. (pp. 11-18), Association for Computational Linguistics.

The author acknowledges the popularity of lexicalized approaches to grammar (such as link grammar). She points out that while other approaches to grammar such as dividing words into word classes saves time, lexicalized grammars provide information on the distinctive nature of words. As she points out, this can also be a challenge because lexicalized grammars contain so much specific information. In this paper, she uses a lexicalized grammar approach (Link grammar) for German. The author further describes Link grammar in the following section. In essence, Link grammar is lexicalized, context-free, non-hierarchical and focuses on the connections between words in a sentence. The words are linked together by different labeled arcs (links), which cannot cross each other, and all words in the sentence must connect. If these requirements are not filled, as the author points out, the sentence is not in compliance with the language the grammar models. One advantage the author points out in this section is that there currently exists a parsing algorithm (Sleator & Temperley, 1991). The author then details adaptations to link grammar to cover the German language. The author outlines advantages of link grammar for learning. The author lists its non-hierarchical nature (links can be learned independently and examined individually), and an absence of long-distance dependencies. This author outlines several advantages to using a Link grammar system, which is similar to a probabilistic approach in that it is also non-hierarchical, which provides reason to use a non-hierarchical model to examine language acquisition in children in my thesis.

Manning, C. D. (2003). Probabilistic syntax. In S. Jannedy, J. Hay, & R. Bod (Eds.) *Probabilistic linguistics* (pp. 289-341). Cambridge, MA: Bradford.

The author provides some background into the research done on probabilistic linguistics. It was an unpopular field of study for several decades, but has recently become an increasingly popular approach to linguistics. The author provided one particularly compelling reason for a probabilistic approach: “Human cognition has a probabilistic nature: we continually have to reason from incomplete and uncertain information about the world, and probabilities give us a

well-founded tool for doing this," (p. 290). The author outlines some of the advantages and some of the drawbacks of using a corpus to study linguistics. He points out that a corpus is useful when searching for lexicon, but is more difficult to analyze syntactic phenomena. The author uses the specific example of verbal clausal subcategorization frames to examine the problems with categorical models of syntax, as well as to demonstrate a probabilistic model and examine its issues. In examining the problems with categorical models, the author points out "language is used more flexibly than such a model suggests" (p.298). When examining subcategorization under a probabilistic light, the author concludes "such models combine formal linguistic theories and quantitative data about language use in a scientifically precise way," (p. 304). In both cases (either a categorical or probabilistic model), the author points out "there is clearly a trade-off between simplicity of the theory and factual accuracy," (p. 306). However, he goes on to mention that probabilistic approaches can help deal with the complexities of language and help make the process more manageable. The author of this chapter provides a compelling argument for approaching syntax with a probabilistic model rather than a categorical model. In my thesis, I aim to examine children's language through a probabilistic model. The rationale for probabilistic models gives motivation for my thesis, as children's language has not yet been widely examined using a probabilistic model.

Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition, 90*, 91-117. doi:10.1016/S0010-2077(03)00140-9

The author introduces the idea that frequent frames, or "distributional patterns based on co-occurrence patterns of words in sentences," can give insight into childhood acquisition of grammatical categories (p.91). Using corpora from the CHILDES database, Mintz conducted two experiments. In the first one, he used the 45 most frequent frames of each language sample to categorize words. In this experiment, he found that frequent frames were an effective method for categorizing words. This experiment yielded high accuracy. In the second experiment, he analyzed the frequency relative to the total number of frames in a sample. This experiment was also done to ensure that the high accuracy obtained in the first experiment was not due to a small number of categories. Based on information from both experiments, Mintz found that frequent frames were an effective method for producing accurate categories. Mintz offers a unique

perspective in this study that frequent frames algorithm can give insight into children's acquisition of grammatical categories. My thesis similarly uses a computerized model to offer additional insight into acquisition of grammatical categories in children.

Redington, M., Chater, N., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science, 22*, 425-469. doi: 10.1207/s15516709cog2204_2

The authors introduce the lack of research in computational approaches to children's language acquisition, particularly how children acquire syntactic categories. This study examines how distributional information plays a role in this acquisition. They mention how this is difficult to explain from both nativist and empiricist perspectives, thus they seek for information from environmental input to explain the children's acquisition of grammatical categories. The authors explain the perspective of distributional information, or the "linguistic contexts in which a word occurs," (p. 427). They also concede that while this won't explain acquisition of GWCs in its entirety, distributional information can provide more insight into the process. Other approaches are presented that had been proposed to give similar insight including semantic bootstrapping, phonological constraints, prosodic information and innate knowledge. Ultimately, the authors want to look at the input that the child receives from their environment to give insight into their language acquisition process. Taking adult language samples from the CHILDES database, Redington, Chater and Finch assigned target words (the 1,000 most frequent words) and context words (the 150 most frequent words) to their most common syntactic category (constraining each word to only one possible category). These results were scored by accuracy, completeness and informativeness. Their results found that the nearer a context word was to a target word, the more it informed about the word's grammatical category. They found that preceding context words were more informative that succeeding context, but the best results were found by combining the two. The findings of their experiments were consistent with their hypothesis that distributional information gives insight into children's acquisition of syntactic categories. Their findings most closely side with an empiricist approach to children's acquisition of language. Similar to the computational approach in this study, my thesis will be examining children's acquisition of language based on where words occur in relation to each other. Rather than study

the distributional information as was done in this study, my thesis examines the probabilities that one grammatical tag will follow another tag in the training corpus and then uses those probabilities to organize a randomly ordered set of grammatical tags taken from the test corpus.

Romberg, A. R., & Saffran, J. R. (2010). Statistical learning and language acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*, 906-914.

Statistical learning refers to recognizing patterns from an input. The field of child language acquisition has recently taken interest in statistical learning because of the fast rate at which children are able to obtain the complex structures of language. Statistical learning in language deals mainly with transitional probabilities, or the probability that a certain structure will follow a certain structure in a sentence. This field of study is expanding in different areas. First, by applying statistical learning to different levels of language (phonemes, words, phrase level, etc.). Second, to connect language acquisition with other cognitive mechanisms. Finally, to determine if statistical approaches are valid in the context of complex, natural language. As the field of statistical language learning expands in these areas, a greater understanding of its extent and limits will provide insight into the role it plays in child language acquisition. My thesis seeks to add to this field of statistical learning of language as the authors call for and uses transitional probabilities, which the authors outline in this article.

Saffran, J. R. (2003). Statistical language learning mechanisms and constraints. *Current directions in psychological science*, *12*(4), 110-114.

Children are able to acquire language, which is a highly complex system. It is thus likely that the mechanism underlying such an acquisition is complex as well, which is why much research has been done in the area of child language acquisition. Two schools of thought presented in this article are theories where learning is central (behaviorist) and theories where learning plays only a small role (nativist). Learning-oriented theories have a lot of validity in that a wide base of research suggests that children have powerful learning mechanisms. These theories are not without their weaknesses, however. Learning-oriented theories don't take into account cross-linguistic similarities, which are better explained by a nativist theory where children are born with a preexisting knowledge of language. The author suggests that a better

understanding of constrained learning will lead to linkages between both behaviorist and nativist theories. One particular study the author conducted to research learning mechanisms involved in language was to examine word segmentation by exposing adults, first-graders and 8-month olds to nonsense language where the only word boundary cues were statistically derived. Their results found that even the infants used statistics to determine word boundaries. The author points out that the area of language most affected by the nativist vs. behaviorist approach is syntax, which the current study examines. The author concludes that statistical cues do help learners in language acquisition, but the extent to which this is true remains to be discovered through future research. Similarly, my thesis hopes to contribute to the line of research regarding statistical learning of language.

Sleator, D. D., & Temperly, D. (1993). Parsing English with a link grammar. *Third International Workshop on Parsing Technologies.* Tilburg, Netherlands and Durbuy, Belgium, 1-14.

This paper defines and outlines link grammar as a grammatical system. The general concepts underlying link grammar include planarity, connectivity and satisfaction. Planarity means that the links connecting the words cannot cross other links. Connectivity means that all of the words in the sentence are connected by links. Satisfaction means that all of the links satisfy the link requirements of the words in the sentence. These linking requirements are contained in a dictionary as outlined by Sleator and Temperly. In link grammar, words are given specific "connectors" that need to be satisfied by other connectors from other types of words. The rest of the paper is divided into 6 more sections. In section 2, Sleator and Temperly define link grammars on a more specific level and delineate the specific notations and terminology found in link grammar. In section 3, an example of link grammar for English is examined. In section 4, the algorithm used in link grammar is described. In section 5, the data structures used to make the program run fast are described. In section 6, the relationship between link grammars, dependency syntax and categorical grammars is examined. Finally, in section 7, other research endeavors that use link grammar are mentioned. This paper delineates the link-grammar model, which is similar to a probabilistic approach in that it is another alternative to hierarchical grammar. My thesis seeks to examine children's language from a similar non-hierarchical model.

Szmrecsanyi, B. 2004. On operationalizing syntactic complexity. In *Le poids des mots. Proceedings of the 7th International Conference on Textual Data Statistical Analysis. Louvain-la-Neuve, March 10–12, 2004*, ed. by G. Purnelle, C. Fairon, and A. Dister. Louvain-la-Neuve: Presses Universitaires de Louvain, 1032–39.

Syntactic complexity has received recent attention in the literature, yet when used as a variable in research, it is underdefined. In order to rectify this problem, the author of this article examines three aspects of syntactic complexity—node counts, word counts, and index of syntactic complexity—and compares them with regard to accuracy and applicability. To compare the differences of these three approaches, all three methods were applied to two different corpora. The author found that all three approaches essentially measure the same thing. The author proposes that word count, which is the most efficient method of the three, should be the preferred method in measuring syntactic complexity as it is as accurate and applicable as the other two. This is relevant to my thesis as it deals with how syntactic complexity is assisted by statistical information. The author of this article is quantifying it on the other end—how to characterize syntactic complexity once its organized.

St. Clair, M., Monaghan, P. & Christiansen, M. H. (2014). Acquisition of grammatical categories. In P. Brooks & V. Kempe (Eds.), *Encyclopedia of language development* (pp. 253-255). Thousand Oaks, CA: Sage Publications.

The authors of this article outline different theories behind children's acquisition of grammatical categories. They introduce nativist and empiricist approaches. The nativist approach argues that children map new words to pre-existing grammatical categories while the empiricist approach argues that children obtain these categories from extracting information from language input through learning mechanisms. The authors delve into distributional contexts, which align with the empiricist approach. Originally proposed by Charles C. Fries in 1952, distributional context looks at words that frequently occur together to extract grammar. This theory was unable to be tested until later when advances in technology allowed for better data collection and investigations. Researchers were later able to provide validity to this approach, finding that frequent frames are informative and help children to acquire grammatical categories. The authors

then outline other cues that could contribute to grammatical acquisition. These include phonological cues, gestural, attentional and semantic cues. The authors concede that while a full picture of how acquisition of grammatical categories is achieved is not clear, research has offered several pieces to the puzzle. They call for future research in this field to find new ways to examine various cues to explain grammar acquisition, as well as determine if an innate structure is involved. In the current study, we hope to add another piece to this puzzle of how children acquire grammatical categories like the authors provided a case for in future research. The current study examines how children use statistical structure as cues for GWC acquisition.

Stenquist, N. A. (2015). *Modeling children's acquisition of grammatical word categories from adult input using an adaptation and selection algorithm.* (unpublished master's thesis). Brigham Young University, Provo, UT.

The author provides insight into nativist and constructivist theories regarding children's acquisition of grammatical categories. The nativist theory argues that children are innately born with these categories. The constructivist theory proposes that children are not born with these categories, but rather extract them from the language they hear. Both theories assume that children have a sophisticated processing mechanism in order to complete bootstrapping, or making sense of the language input they are exposed to. The author outlines previous research that has been performed that has examined syntactic bootstrapping in order to address grammatical word category (GWC) acquisition. The author then offers the perspective of an evolutionary model to address GWC acquisition. The author examined both adult and children's utterances in the corpora, the adult corpora used for training the computer model and the child corpora for testing the model. The computer model used in this study used an adaptation and selection model to evolve a set of GWCs given exposure to language input. The study found a rapid increase in accuracy in generations 1-1500, and a gradual increase in accuracy until generation 12,000 in all of the children's language samples. The study found that using an evolutionary model was highly successful in achieving accuracy in assigning words to grammatical categories. The author compared results to previous studies, and also offered unique contributions from the present study. One such example was that the present study evaluated parent's language from their particular children's language. The present study also examined a

larger amount of language than previous studies. The study provides validity to an adaptation and selection algorithm in examining children's acquisition of grammatical categories. The current study will be using a probabilistic computer model to examine the role of probability in children's acquisition of grammatical categories. Stenquist's study provided perspective to children's acquisition of grammatical categories using an evolutionary computer model, and my thesis hopes to provide further perspective by using a probabilistic model.

Stumper, B., Bannard, C., Lieven, E., & Tomasello, M. (2011). "Frequent frames" in German child-directed speech: A limited cue to grammatical categories. *Cognitive Science, 35*, 1190-1205. doi: 10.1111/j.1551-6709.2011.01187.x

This study builds off of Mintz's 2003 study of examining how frequent frames give insight into childhood acquisition of grammatical categories. While Mintz's study found that frequent frames were an effective method to producing accurate categories in English, this study by Stumper et al. seeks to see if the same is true for German. An earlier study by Chemla et al. (2009) extended frequent frames to French. As German has a less restricted word order than French or English, this study wanted to examine if accuracy and completeness would remain high as it did in the other two studies. Using German corpora from the CHILDES database, the authors performed the same study that Mintz performed in 2003. The results found that frequent frames in German were less accurate than in English or French. The authors concluded that children have to probabilistically use many sources from their input to acquire grammar. As with this study, the current study uses a computerized model to offer additional insight into grammar acquisition. As my thesis only examines English, future research could examine the role of statistical structure in grammar acquisition of other languages as well.

Wintner, S. (2010). Computational models of language acquisition. In A. Gelbukh (Ed.), *Computational linguistics and intelligent text processing* (pp. 86-99). Berlin, Germany: Springer. doi: 10.1007/978-3-642-12116-6_8

In this paper, the author examines computational models of child language acquisition from fields of psycholinguistics, cognitive science and computer science. Computational models give valuable insight into the language acquisition process, and combining approaches from all

different fields can provide a more holistic picture to this phenomenon. Like children, computer programs are “learners” in that they are presented with data, draw information from the data to then generalize. Computational models are distinguished by the data, task, grammar and evaluation. In evaluation of computer learning models, several difficulties exist. For example, the training data, while extensive, is still much less than what the child is exposed to overall. A measure proposed by Chang et al. seeks to address the problem of evaluation. Sentence prediction accuracy refers to a learner’s ability to reorder the words of a sentence when presented in a scrambled order. The current paper uses the same measure to evaluate the computer model. The author then outlines current computer model approaches and offers direction for future research. Namely, the fields of psycholinguistics, cognitive science and computer science, which have previously worked as separate entities, should come together to better examine child language acquisition. Overall, the author provides a convincing argument for the use of computer models in examining (child language acquisition, which my thesis uses.