Spring 2009

A neurophysiological study on probabilistic grammatical learning and sentence processing

Hsin-jen Hsu
University of Iowa

A NEUROPHYSIOLOGICAL STUDY ON PROBABILISTIC GRAMMATICAL LEARNING AND SENTENCE PROCESSING

by

Hsin-jen Hsu

An Abstract

Of a thesis submitted in partial fulfillment
of the requirements for the Doctor of
Philosophy degree in Speech and Hearing Science
in the Graduate College of
The University of Iowa

May 2009

Thesis Supervisor: Professor J. Bruce Tomblin

ABSTRACT

Syntactic anomalies reliably elicit P600 effects in natural language processing. A survey of previous work converged on a conclusion that the mean amplitude of the P600 seems to be associated with the goodness of fit of a target word with expectation generated based on already unfolded materials. Based on this characteristic of the P600 effects, the current study aimed to look for evidence indicating the influence of input statistics in shaping grammatical knowledge/representations, and as a result leading to probabilistically-based competition/expectation generation processes of online sentence processing. An artificial grammar learning (AGL) task with 4 different conditions varying in probabilities were used to test this hypothesis. Results from this task indicated graded mean amplitude of the P600 effects across conditions, and the pattern of gradience is consistent with the variation of the input statistics. The use of the artificial language to simulate natural language learning process was further justified with statistically undistinguishable P600 effects elicited in a natural language sentence processing (NLSP) task. Together, the results indicate that the same neural mechanisms are recruited for both syntactic processing of natural language stimuli and sentence strings in an artificial language.

Abstract Approved: _____
Thesis Supervisor

_____
Title and Department

_____
Date

A NEUROPHYSIOLOGICAL STUDY ON PROBABILISTIC GRAMMATICAL LEARNING AND SENTENCE PROCESSING

by

Hsin-jen Hsu

A thesis submitted in partial fulfillment
of the requirements for the Doctor of
Philosophy degree in Speech and Hearing Science
in the Graduate College of
The University of Iowa

May 2009

Thesis Supervisor: Professor J. Bruce Tomblin

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

_____

PH.D. THESIS

_____

This is to certify that the Ph.D. thesis of

Hsin-jen Hsu

has been approved by the Examining Committee
for the thesis requirement for the Doctor of Philosophy
degree in Speech and Hearing Science at the May 2009 graduation.

Thesis Committee: _____
J. Bruce Tomblin, Thesis Supervisor

_____
Karla K McGregor

_____
Bob McMurray

_____
Toby J Mordkoff

_____
Amanda J Owen

ACKNOWLEDGEMENTS

On completion of this dissertation, I would like to gratefully and sincerely thank my advisor Dr. Bruce Tomblin for his support, encouragement, and mentorship over the past four and half years. I cannot remember a single time I brought up a research idea to him and I was not encouraged by his positive response. Bruce first led me into this area of research in my pre-dissertation research project, which later motivated this thesis. Without his unique “doable” spirit in research and all the assistance he provided me, this thesis is impossible.

I would like to acknowledge the members of my Thesis Committee. Particular, I would like to thank Dr. Toby MordKoff for his time and knowledge in ERP techniques. I appreciated the time Toby spent sitting down and going through my experiments and questions with me. During the stage of data analyses, his suggestions and comments have greatly benefited me. I would also like to thank Dr. Amanda Owen, Dr. Karla McGregor, and Dr. Bob McMurray for their knowledge, thought, and patience during the revision of this thesis. Their feedback and comments in reading an early draft of my thesis and the questions they raised in my defense have helped me to think the questions I raised in my thesis more deeply. I am grateful to Amanda for her mentorship during my doctoral study and her help with my thesis writing. I appreciated that Amanda’s door was almost always open, and when it wasn’t she would get back to me.

I would like to thank Dr. Morten Christiansen for taking time to discuss my thesis with me on his visit to Iowa. I benefited greatly from his knowledge and experience in research. I would also like to thank Dr. Christine Weber-Fox and Dr. Patricia Zebrowski for letting me use their devices and lab space to carry out my experiments. I also greatly indebted to Wendy Fick, who helped me to get things started in the stuttering lab, and Rick Arena, for not only being a good friend but always a great help with task

development. A special thank goes to Marlea O'Brien, Amanda Burns, Victoria Tumanova, Derek Stiles and Nichole Eden, for being supportive friends and being generous with their time helping me record my stimuli.

I would like to give a very special thanks to Lu Chunming, Zao Libo, and Joe Toscano for their knowledge and assistance with data analysis. I would like to thank Michelle Quinn for reading an early draft my thesis and provided me valuable suggestions. I am also in indebted to Linda Spencer, who has been very supportive and helpful with my study.

I would like to give very special thanks to Vicki Samelson for being not only a supportive colleague, but also a friend through the ups and downs of research. A very special thank also goes to Hugo Ling-yu Guo, for being a good friend and for helping me to deal with some paper work and send my thesis in when I was not able to be there.

Finally, my most heartfelt gratitude and undivided love go to my fiancé Ting, a kind and loving man who always put me in the first place. Thank you for patiently listening to me going on about every detail in my thesis so that I can better organize my thought. You put down your work and flew from UK to make sure I eat and sleep properly while I was working days and nights during data collection. I am not sure if I would have survived the final stages of thesis without you.

ABSTRACT

Syntactic anomalies reliably elicit P600 effects in natural language processing. A survey of previous work converged on a conclusion that the mean amplitude of the P600 seems to be associated with the goodness of fit of a target word with expectation generated based on already unfolded materials. Based on this characteristic of the P600 effects, the current study aimed to look for evidence indicating the influence of input statistics in shaping grammatical knowledge/representations, and as a result leading to probabilistically-based competition/expectation generation processes of online sentence processing. An artificial grammar learning (AGL) task with 4 different conditions varying in probabilities were used to test this hypothesis. Results from this task indicated graded mean amplitude of the P600 effects across conditions, and the pattern of gradience is consistent with the variation of the input statistics. The use of the artificial language to simulate natural language learning process was further justified with statistically undistinguishable P600 effects elicited in a natural language sentence processing (NLSP) task. Together, the results indicate that the same neural mechanisms are recruited for both syntactic processing of natural language stimuli and sentence strings in an artificial language.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

CHAPTER I

INTRODUCTION

1.1 Statistical Language Learning

To what extent is language, or more specifically grammar, learnable from experience given the remarkable speed with which human infants acquire a language in their first years of life? This question is the basis for the "poverty of stimulus" (POS) argument (Gold, 1967; Chomsky, 1965; Pinker, 1984). Specifically, the POS argument stated that positive evidence available to language learners is insufficient to allow them to distinguish grammatical from ungrammatical utterances. Because language input seems degenerate, in that speech is full of flaws and missing elements as well as lacking negative evidence, successful learning was hypothesized to occur on a deductive basis by means of an innate language learning device.

Recently, there has been a renewed interest in considering language acquisition from an inductive perspective. This recent and renewed interest is contributable to accumulated evidence that the language input children hear contains abundant statistical regularities (Kelly 1992; Kelly & Martin, 1994; Billamn & Knutsen, 1996). In addition, recent studies have shown that children are equipped with strong data mining abilities enabling them to detect distributional regularities in the input and learn linguistic categories and structures at various levels (Maye, Werker, & Gerken, 2002; Saffran, Newport, & Aslin, 1996; Saffran, Aslin, & Newport, 1996; Saffran, Johnson, Aslin, & Newport, 1999; Gómez & Gerken, 1999; Altmann, 2002b; Gerken, Wilson, Lewis, 2005). We can broadly refer to this field as statistical language learning, and this term is used hereafter in this thesis. By using this term we do not wish to go beyond the idea of a

critical role of statistics in language acquisition to suggest that statistical learning is language specific. In fact, the domain generality of statistical learning has been well documented (e.g., Fisher & Aslin, 2002; Conway & Christiansen, 2005; 2006), and is generally conceived of as one of the general-purpose mechanisms that broadly support human learning.

The discoveries of the rich array of statistically structured input available to the child and the strong learning capacities that children bring to the task of language acquisition have countered the dominant linguistically motivated theory that about whether language acquisition ultimately necessitates a language-specific device (LAD) such as a Universal Grammar. In particular, studies using miniature language or grammars have successfully shown that young infants and adults are capable of utilizing input statistics to detect underlying structures, suggesting inductive basis of grammatical acquisition.

Given this recent development; however, debate continues on the two general approaches to language acquisition especially in the area of syntax (Bonatti, Peña, Nespor, Mehler, 2005; Gómez & Gerken, 1999; Marcus, 1999; 2001; Marcus & Berent, 2003; Marcus, Vijayan, Bandi Rao, & Vishton,1999; Keidel, Jenison, Kluender, & Seidenberg, 2007; Onnis, Monaghan, Christiansen, & Chater, 2004; Onnis, Monaghan, Richmond, & Cater, 2005; Seidenberg, MacDonald, & Saffran, 2002; 2003). For instance, using a finite-state grammar, Gómez and Gerken (1999) reported successful generalization of sequential word order in a group of 12-month-olds and concluded that their infants are capable of making use of statistical regularities in the input to discover and generalize sequential structures. Similar results from younger 7-month-olds were obtained by Marcus, Vijayan, Bandi Rao, and Vishton (1999). However, Marcus and colleagues interpreted their findings as evidence that the infants learned algebra-like rules that operate on abstract unitary symbols. Marcus (2001, 2003) further argued that the same interpretation is applicable to the data obtained by Gómez and Gerken (1999), as

people can learn formal patterns that hold any element, irrespective of its statistical properties.

More recently, Peña, Bonatti, Nespor, Mehler (2002) voiced a compromise position that certain aspects of language acquisition, such as sound categories, speech segmentation and vocabulary learning, operate on the basis of statistical learning, whereas entirely separate algebraic computations are necessary for learning grammatical structures. Although Peña and colleagues' proposal seems to have provided basis for reconciling the issue, their proposal in fact is not different from Marcus and colleagues' argument at least in terms of syntactic learning. As would be expected, this view does not settle the argument, but instead has stimulated a series of debates between groups of researchers (Bonatti et al., 2005; Onnis et al., 2005; Keidel et al., 2007). This debate is well captured by the title "*Does grammar start where statistics stop*" p.553) in the paper by Seidenberg, MacDonald, and Saffran (2002).

Thus, at this time a central issue in grammar acquisition concerns the nature of what is learned and processed during development: rules or statistics. This issue constituted the focus of this thesis. In this thesis, we addressed this issue by collecting complementary data from a natural language sentence processing (NLSP) task and an artificial grammar learning (AGL) task. Those who argued for abstract rules assume that this knowledge arises from language specific mechanisms that operate on only natural language input. Thus, the most convincing data comes from a natural language sentence processing (NLSP) task in which statistical learning play a critical role. To examine whether statistical learning processes can yield a rule-like product, we need to carefully control the properties of the language. Thus we also use artificial language structure designed to produce stochastic like knowledge about sequences in an AGL task. An AGL task has the added advantage that performance can be measured at the beginning of learning a grammar and at a latter point of learning mastery. Artificial language learning tasks including statistical learning tasks are always open to questions as to whether these

actually engage the mechanisms used in natural language. Thus, comparisons of performance in the two tasks constituted the third major primary question in this thesis.

In the current study, event-related potential (ERP) data was collected to compare syntactic processing of sentences in the AGL and the NLSP task. Within the area of syntax, it is known that syntactic incongruity can elicit a P600. This is an ERP component that is often termed Syntactic Positive Shift (SPS). This terminology reflects the fact that earlier studies focusing on the P600 component were conducted exclusively in language studies. A wide variety of grammatical violations have been reported to elicit the P600, including morphological violations of number, gender, and case (Atchley, Rice, Betz, Kwasny, Sereno, & Jongman, 2006; Coulson, King, & Kutas, 1998; Osterhout & Mobley, 1995), phrase structure (Hagoort, Brown, & Groothusen1993; Neville, Nicol, Barss, Forster, & Garrett, 1991), subjacency (McKinnon & Osterhout 1996; Neville et al., 1991), and subcategorization (Ainsworth-Darnell, Shulman, & Boland, 1998; Osterhout & Holcomb, 1992). More recently, new findings have shown the P600 component is not specific to language. Instead, this component can be elicited by non-linguistic materials. Furthermore, the magnitude of the P600 seems to vary with the degree of deviation from expectations (e.g., Patel 1998; 2003; Patel, Gibson, Ratner, Besson, & Holcomb, 1998). Magnitude variation in particular is informative for the question of rules versus statistics. Statistical representations allow for graded expectations, whereas algebraic rules have considerable difficulty with such graded expectations. An AGL task that has different probabilistic structures built in should therefore provide graded P600 responses associated with deviation from expectations based on their probabilities.

If violations of rules learned from an AGL produce the same ERP features as those generated by violations of natural grammar, including patterns reflective of graded probabilities we should have evidence in support of a common mechanism that tracks and stores statistical information. If these patterns for natural language are different from those of the AGL and, in particular, not graded, the evidence would suggest that a

statistical account of natural language processing is incorrect and that grammatical rules may more accurately capture the profile of performance.

1.2 Specific Aims of the Study

In the previous section, we argued that a dominant issue within psychology of language concerns the nature of grammatical knowledge and the mechanisms that give rise to this knowledge. The dominant paradigm has been one in which algebra-like rules operate on unitary symbols to account for the sequence patterns of sentences in natural language. An alternative account argues that these sequences are formed from statistics provided in the input. Using converging evidence based on ERP analyses of P600 derived from natural and artificial language tasks. We addressed the following questions:

I. Does the P600 reflect probabilistic (expectancy) knowledge of sequences? This question was addressed more specifically via the following questions.

1. Within the AGL task, is the magnitude of the P600 response associated with the developmental status of the learner such that no response is seen at the beginning of learning and a greater response is found at the end when behavioral measures show high levels of grammatical knowledge?

2. After training in the AGL task is the magnitude of the P600 associated with the conditional probability of the “rules”?

II. Within the NLSP task in which the arguments of transitive and intransitive verbs are varied, does the magnitude of the P600 vary in accord with the violation of likelihood that the verb would take a particular argument?

III. Does the P600 obtained in the AGL task conform to the P600 obtained in the natural language task?

1.3 Organization of this Thesis

First, in chapter 2, we will address the developmental questions that have shaped the current debate on grammatical acquisition. A distinction will be made between the perspective on language acquisition within the traditional generative/minimalist program and how the central questions and premises of this theoretical paradigm can be explained by the statistical language learning approach. This will be followed by a summary of empirical evidence supporting statistics as useful cues for the discovery of categories and structures of language at various levels.

Having laid out the distinction between the two approaches and the extent to which statistical learning can achieve language acquisition, we will turn to discuss implications of the two approaches for sentence processing and how they differ in terms of expectancy generation and the state of representation activations when processing sentences.

Following the discussion, we will further address the current debate between the statistical learning and innate, rule-governed nature of language acquisition. We will argue that one difficulty in teasing apart the two accounts has to do with working back from data to the underlying processes that give rise to it, and this can be solved by using measures that are capable of monitoring performance on-line. This leads to a discussion about the principal ERP component in this study, P600. We will first review antecedent conditions that have been reported to elicit the P600, a syntactic-related ERP component, from earlier studies and then show how results from more recent studies have converged on domain-general, learning-based, and statistically-based expectation generation nature of the P600. Particularly, we will focus on studies by Patel and colleagues' finding of statistically indistinguishable P600 effects elicited by linguistic and musical materials.

Additionally, studies from musical perception and second language acquisition will also be reviewed to show the relationship between the P600 component and experience. The work by Patel and colleagues is valuable in re-thinking and re-evaluating earlier language studies on the P600 such as that by Osterhout, Holcomb, and Swinney (1994). In this context, we will show how data from their study provides support for statistical learning of grammar. Some concerns regarding the task used in their study will also be raised.

A detailed description of the tasks used in the current study is provided in chapter 3. This Chapter will start with a description of the AGL task used in the current study. In the language, words from the same "category" vary in terms of the probabilities of co-occurrence with a certain subsequent elements. A replication of one sentence processing task used in the 1994 study by Osterhout and colleagues constituted the second principal task of this thesis. Following the concerns raised in Chapter 2, we will discuss three major modifications we made to form the current version of the task used in this thesis. Results from a multi-corpus analysis, which was conducted to confirm the verb categorizations in the 1994 study by Osterhout and colleagues, were also reported. In Chapter 4, we will report results from the behavioral measure of grammaticality judgments and ERP responses in the AGL and the NLSP tasks. Additional analyses and discussion on individual differences in learning the artificial grammar, the results from the corpus analysis, and overall comparisons of the AGL and NLSP tasks will be reported in Chapter 5. Finally, a discussion of the results from the current findings and a conclusion will be provided in Chapter 6.

CHAPTER II

BACKGROUND AND THE PRESENT RESEARCH

2.1 Statistical Language Learning

The debate on how children acquire a language has a long history in developmental psychology. Earlier in the 20th century the idea that children might learn by means of associative principles was widespread (Skinner, 1957). It was believed that language, like other behavior, is subject to the rules of operant conditioning, which are based on associative principles. Learning verbal behavior or language under this framework was therefore conceptualized as a process of building stimulus-response associations and strengthening and maintaining the association via reinforcement, a process guided by the *probability of occurrence* between a stimulus and a response in the input (Palmer, 1981).

The early associationist work was discounted with the emergence of the generative paradigm which argued that associative mechanisms were inadequate for acquiring the complex structure of human language (Chomsky, 1959). First of all, in the traditional generative/minimalist paradigm language acquisition is assumed to be a process of setting innate, domain-specific linguistic parameters which required only a limited amount of language input. The “poverty of the stimulus” argument claims that the input to the child is degenerate, comprised of both grammatical and ungrammatical sentences that are not labeled as such (i.e., a lack of negative evidence) thus grammar cannot be induced from the input provided to the child. However, children learn language over a short span of time with such seeming ease even though they have no reliable evidence from language input. It was therefore assumed that the only possible solution to

the induction problem and explanation for how language can be learned rapidly under unsupervised conditions is that it must be a kind of human instinct and that children learn language by means of a language-acquisition device (LAD, Chomsky, 1965; Wexler & Culicover, 1980). Such a device is thought to narrow down the search space of possible languages by limiting the parameters children search for in language (Hyams, 1986; Manzini & Wexler, 1987; Roeper & Williams, 1987). As a result, the role of input is merely to allow children to learn a lexicon and set linguistic parameters. Children at young ages are considered to begin to process language using rules that operate on abstract syntactic categories or symbols, such as subject noun, object noun, and verb.

More recently, however, domain-general learning approaches, such as a statistical learning account, have been gaining credibility, spurred by discoveries of abundant regularities in the language input the child hears (Billamn, 1989; Kelly & Bock 1988; Kelly 1992; Kelly & Martin, 1994; Billamn & Knutsen, 1996). For instance, Kelly and Bock (1988) analyzed stress patterns of English disyllabic nouns and verbs using a sample containing over 3000 nouns and 1000 verbs and found that 94% of words with first-syllable stress were nouns rather than verbs, whereas 85% of words with second-syllable stress were verbs rather than nouns. In addition to stress, other phonological cues such as syllable number, word duration, vowel epenthesis, and voicing all probabilistically correlate with grammatical class (Kelly, 1992). The existence of statistical regularities at multiple linguistic levels has been shown to be of use in acquiring language. Studies during the last 10 years have shown that young infants, children, and adults are capable of exploiting statistical information such as raw frequency, frequency of co-occurrence, or transitional probability to learn speech categories (Anderson, Morgan, & White; 2003; Maye et al., 2002; Maye & Weiss, 2003), to detect phonotactic structures (Chambers, Onishi, & Fisher, 2003), to locate word boundaries in running speech (Saffran et al., 1996; Saffran, 2001a), to learn word sequences (Gómez & Gerken, 1999), to form syntactic categories (Altmann, 2002 b;

Gerken et al., 2005), to track long-distance dependencies (Gómez, 2002; Onnis et al., 2004), and to build abstract representations of sequential patterns (Gerken et al., 2005; Gómez & Gerken, 1999; Gómez & Lakusta, 2004). The robustness of the statistical learning abilities was demonstrated by successful learning after only a brief exposure to the language. These findings together have posed a direct challenge to the generative/minimalist paradigm (Chomsky, 1965; 1995) since statistical learning or any account which claims a fundamental role to segmentation, categorization, and generalization is rejected in Chomskyan linguistics as "mistaken in principle" (Chomsky, 1957). In addition, the finding that language learners are capable of using their data mining abilities to detect distributional regularities in the input to acquire language structures has led to the rejection of the "poverty of the stimulus" argument. These studies also suggest that stochastic languages may be learnable from positive examples alone. This is because the structural regularities of language can be derived from relatively noisy input data, while their non-stochastic analogues require negative evidence (Gold, 1967; Angluin, 1988).

Despite the success of the statistical language learning, there have been questions raised about whether such mechanisms can succeed in learning complex structures that are not tied to the surface properties of the input, such as hierarchical phrase structures. Phrase structure refers to groupings of words into sub-units (i.e., constituents) which may then themselves become a new sub-unit and therefore resulting in hierarchically organized groups of elements. For instance, the words in "*The space probe sent back images of Mars*" fall into particular groupings (*The* (*space probe*)) (*sent back* (*images of Mars*)) rather than random (e.g., (*The space*) (*probe sent back*) (*images of*) (*Mars*)). To examine if statistical learning could learn hierarchical phrase structure, Saffran (2001b, 2002) compared learning of two artificial languages, one containing predictive dependencies between words and the other lacking predictive dependencies. In the predictive language the presence of a word token belonging to a word category was

always preceded by the occurrence of a word belonging to a different word category. On the other hand, in the non-predictive language the relationship between the two categories is variable. She found that both adult and children learners exposed to the language containing predictive dependencies performed better in detecting phrasal units than learners exposed to the language lacking predictive dependencies.

How does statistical learning of predictive dependencies inform us about language use? One intriguing implication from Saffran's work in particular, as well as the statistical language account in general, is that the acquired statistical relations not only provide basis for expectancy generation, a phenomenon that has long been recognized as key in processing words and comprehending sentences (Kutas & Hillyard, 1984; Marslen-Wilson, & Welsh, 1978; Altmann, 2002a; Federmeier & Kutas, 2001; van Berkum, Brown, Zwitserlood, Kooijman, & Hagoort, 2005; Wicha, Moreno, & Kutas, 2003), but also imply statistically ranked (and therefore graded) activations of all possible candidates according to the acquired statistics. This idea of expectation generation during on-line sentence processing is consistent with earlier interactive models (Bates & MacWhinney, 1989; Tyler & Marslen-Wilson, 1977; Taraban & McClelland, 1988) or the constraint-based (or constraint-satisfaction) models (e.g., MacDonald, Pearlmutter, & Seidenberg, 1994; McRae, Spivey-Knowlton, & Tanenhaus, 1998; Trueswell, Tanenhaus, & Kello, 1993; Trueswell, Tanehaus, & Garnsey, 1994). In these models, it is generally assumed that the network activates all possible analyses of a sentence in parallel, and that the activation of the analyses depends on the amount of support they receive from various sources of information, including frequency, prosody, discourse contexts, and the effects of probabilities. The assumptions that various sources of information are taken into account during sentence processing provides the opportunity for the acquired statistical information to play a role in shaping the state of activations (competition) such that the degree of activation of possible candidates would vary according to the acquired statistics or probabilities. In other words, expectations are not generated in a random fashion but

are constrained by the probabilities developed over time. Additional support for these models comes from research work on connectionist modeling. For example, error-based learning algorithms, such as back-propagation, in connectionist networks use the difference between an anticipated output and the target output in order to adjust the weights that were responsible for the prediction (Rumelhart, Hinton, & Williams, 1986). One type of back-propagation-trained network, a simple recurrent network (SRN), which predicts the next word at output given the previous word at input, is able to learn syntactic categories and relationships from the sequential structure of their language input (Christiansen & Charter, 1999; 2001; Elman, 1990, 1993; McDonald & Christiansen, 2002). In addition, the SRN has also demonstrated how unsupervised learning could possibly be achieved with prediction, as opposed to innate linguistic knowledge, as one key component in the models.

Interestingly, these interactive models or constraint-based models are broadly in line with the lexicalist accounts of language processing, in that they assume that most or all syntactic information (statistical information of structural regularities) is stored with individual lexical items (MacDonald et al., 1994; MacDonald, 1997, Trueswell, 1996). Under this context, verbs are considered an important source of expectancy generation as they exhibit systematic restrictions on the phrases with which they can or cannot co-occur (its arguments). The idea that verbs are powerful generators of expectancy is supported by empirical studies showing that these restrictions influence comprehension, such as in comprehending ambiguous sentences (for a review, see Tanenhause & Trueswell, 1995). Unlike other categories such as nouns,verb seldom stand alone in speech. Nevertheless, it is worth noting that the suggestion of verbs as strong expectancy generators does not exclude other categories such as nouns and adjectives from exerting constraining forces on the anticipatory process (e.g., McRae, Hare, Elman, & Ferretti, 2005).

The idea that language users anticipate what comes next is not in conflict (at least not directly) with the language acquisition perspective inherited from the

generative/minimalist paradigm or its corresponding modular, two-stage models of sentence comprehension. Expectancy generation can certainly be done by activating subsequent abstract syntactic structures after parsing parts of a sentence. For instance, parsing the first couple words of a sentence into Determiner-Noun could lead to prediction of a subsequent syntactic category of verb. However, the two-stage models contrast greatly with the interactive models or constraint-based models of sentence processing. In two stage models, operations of syntactic cues can be sealed off from the influence of other types of constraints during on-line sentence processing. One of the best known two-stage models is the garden-path model developed by Frazier and colleagues (Frazier, 1990; Frazier & Clifton, 1996). According to this model, a single analysis is chosen based on principles such as minimal attachment defined in terms of a phrase-structure tree in the first stage of sentence comprehension. Minimal attachment stipulates that when there are multiple possible syntactic analyses (i.e., temporary uncertainty or ambiguity), an ambiguous phrase is attached to the preceding tree structure using the fewest number of nodes. The only information used in the first stage in interpreting sentences is coarse-grained abstract syntactic information such as "*she*" is a noun and "*heard*" is a verb. A noun phrase following the verb would always be interpreted as the verb's direct object according to the minimal attachment constraint.

The generative/minimalist paradigm and garden-path processing model contrast greatly with the statistical language learning, interactive, and constraint-based processing models in two ways. First of all, the former group allows for only one interpretation at the initial stage of sentence comprehension whereas the latter group allows for consideration of all possible subsequent structures. More importantly, on hearing a post-verbal noun phrase, the constraint-based models would suggest activations of both the direct object interpretation (*She heard the news*) and the subject noun of a sentential complement clause interpretation (*She heard the news was incorrect*). The activations are ranked according to the probability of occurrence of each of the two constructions (among

others) given the verb "*heard*". Such predictions do not exist in the garden path approaches. Furthermore, even if activations of all possible candidates were allowed, the generative/minimalist paradigm and the garden-path model has no room for the potential influence of acquired statistics in that probabilistic information is excluded from rule-based learning.

2.2 Debates on Grammatical Acquisition

In the previous section, we summarized current findings on statistical language learning and discussed how learning statistics result in a coherent account to the phenomenon of expectancy generation in comprehension, and how statistical language learning contrasts with the traditional generative/minimalist paradigm and it compatible sentence processing models such as the garden-path model. Specifically, we argued that statistical regularities are useful cues for learning linguistic material at various levels, and once statistics are acquired, this information exerts influence on language use. In sentence comprehension, this can give rise to probabilistically ranked (or graded) activations of multiple candidates.

However, one question that is often raised toward the statistical language learning account is whether or not statistical learning is at work in language acquisition. Although there is ample evidence demonstrating the robustness of statistical learning, this does not automatically validate the role of statistics in acquisition and its psychological reality in language comprehension. This, in fact, is the core of the on-going debate between the two accounts discussed in 2.1. In this section, focus was given to issues relevant to teasing apart the two approaches. The discussion in this section shaped the main questions that we attempted to examine in the current study.

Whether or not statistical learning is at work in language acquisition is still an issue open to debate. Central to the debate is the fact that there are parallel statistical and grammatical (for example, learning and using symbolic rules) interpretations of the

results from studies on grammatical learning. This phenomenon is common in the area focusing on grammatical acquisition. For instance, Reber (1969) and Mathews and colleagues (1989) using artificial grammars reported good transfer to letter strings consisting of letters not used in the training stimuli. They concluded that the transfer predominantly relies on abstract knowledge. Similar conclusions have been drawn in Marcus et al. (1999) study in which they exposed 7-month-olds to nonsense strings that contained simple ABA (*wi-di-wi* and *de-li-de*) or ABB (*wi-di-di* and *de-li-li*) word patterns. The underlying pattern was the same for training and for test, however, the vocabulary was different. They found that 7-month-old infants were able to discriminate test strings with the same patterns from those with different patterns despite the change in vocabulary. These results were important for demonstrating that younger infants can also abstract beyond a specific word order. Marcus and colleagues further interpreted these findings as evidence that infants are acquiring algebra-like rules. Marcus argued that systems sensitive only to statistical regularities are, in principle, incapable of such abstraction.

Researchers have argued that generalization based on associative learning does not require an assumption of algebraic-like rules (Brooks and Vokey, 1991; Christiansen & Curtin, 1999; Perruchet, Tyler, Galland, & Peereman, 2004; Vokey & Higham, 2005). In a similar study, Gómez and Gerken (1999) examined this by familiarizing 12-month-olds with a finite-state grammar in one vocabulary and tested them on strings in entirely new vocabulary (infants heard strings like FIM-SOG-FIM-FIM-TUP and were tested on PEL-VOT-PEL-PEL-JIC). Thus, although constraints on word ordering remained the same between training and test, vocabulary did not. Their infants were successful in discriminating strings that followed the patterns from those that violated the patterns. The infant abstraction abilities documented by Marcus et al. (1999) and Gómez and Gerken (1999) are dependent on learning patterns of repeating and alternating elements, e.g. ABB, ABA, ABCA (Gómez , Gerken, & Schvaneveldt, 2000). Gómez and Gerken

(1999) argued that abstraction patterns only required pattern-based abstraction, rather than algebra-like rules.

It is fair to say that Gómez and Gerken's (1999) findings do not show that the conclusions of Reber, Marcus and others are necessarily incorrect, but rather show that there is an alternative interpretation consistent with the data. This situation also reveals one potential difficulty of working back from behavioral data obtained from end-point measures to the underlying processes that give rise to it.

In addition, there have been suggestions about a reconcilist view: it is possible that both statistical processes and grammatical processes involved in language acquisition, with the former contributing to simpler problems such as learning the sounds of a language and identifying words in speech and the latter responsible for discovering higher level (grammatical) structures. In a series of experiments, Peña, Bonatti, Nespor, Mehler (2002) exposed adult learners to a continuous speech stream composed of 3 trisyllabic items characterized by their nonadjacent transitional probabilities for 10 minutes and found that although their participants could identify the words in the stream (Experiment 1), they failed to discover the grammatical-like regularities (Experiment 2), even when the duration of training was extended (Experiment 4). When brief pauses (25 msec) were added at word boundaries, their participants were able to discover underlying structural regularities (Experiment 3), and the same level of performance could be reached with only 2 minutes of exposure to the same materials (Experiment 5). Based on the findings, Peña et al. (2002) concluded that the discovery of components of a stream and the discovery of structural regularities require different kinds of computations; statistical computations are efficient for identifying components of a stream but are insufficient to support discovery of the underlying grammatical-like regularities; therefore, different computations of algebraic or rule-governed nature, as they suggested, are responsible for the observed learning in their Experiment 3 and 5. The results from Peña and colleagues' study were well captured by the question "*Does grammar starts*

where statistics stop?" (Seidenberg, MacDonald, & Saffran, 2002). Since speech segmentation could be achieved by statistical learning, but Peña and colleagues claimed that entirely separate algebraic computations are necessary for learning grammatical structure.

Following the study by Peña et al. (2002), Onnis, Monaghan, and Chater (2005) argued for an alternative explanation: in each experiment, the dependent syllables began with plosives and the intervening syllables began with continuants; it is possible, then, that phonological properties exerted an influence on the results rather than that the participants learned the statistical or algebraic properties of the stimuli. This argument is supported by results from two experiments in which they first replicated successful segmentation in English and then showed that the success in segmentation disappeared after the phonological properties of the stimuli were controlled. Consequently, there is no evidence for learning, either statistical or algebraic, on the basis of the nonadjacent dependencies in the stimuli.

In a subsequent study, Bonatti, Peña, Nespor, and Mehler (2005) used a similar artificial language (also French) but with slightly different dependent relationship between elements than was used in their 2002 study. They presented evidence that they used to claim that statistical learning is guided by innate grammatical knowledge. In the study, adult learners were trained on an artificial language in which words were strings of alternating consonant and vowels (CVCVCV). In the first condition, transitional probabilities between consonants within a word were 1.0, but vowels varied. In a second condition, the transition probability relations between consonants and vowels were inverted: the transition probabilities between vowels within a word were 1.0, but consonants varied. The key finding was that their participants were able to learn in the first condition, but not the second condition. Bonatti and colleagues argued that the results cannot be explained by domain-general statistical learning because, if statistics entirely determined the information content, then equivalent performance should have

been obtained. Bonatti et al. argued that results reveal functional differences between consonants and vowels that are encoded in the language module.

In a commentary, Keidel, Jenison, Kluender, and Seidenberg (2007) reviewed the study by Bonatti and colleague and proposed an alternative, experience-based interpretation to Bonatti et al.'s results. Keidel et al. (2007) conducted a corpus analysis and obtained type and token frequency of three-consonant tiers and three-vowel tiers in 4943 French CVCVCV words and the results were further quantified using information theory. They found that in French, consonants provide substantially more information about lexical identity than vowels, suggesting that the learners in Bonatti et al. (2005) study may in fact attend to consonants not out of any innate constraints, but because a lifetime linguistic experience. However, Keidel et al. (2007) also acknowledged that their results do not provide counter evidence to the conclusions of Bonatti et al., but rather show that there is an alternative interpretation can also explain their data. This, again, is the same issue mentioned earlier between Gómez and Gerken's (1999) argument and that of Reber and Marcus and colleagues'.

What is the significance of claiming that statistics play a role in language acquisition? Specifically for the current study, the statistical information acquired in language acquisition is intimately associated with language use, such as expectancy generation in comprehension. This is one of the implications of the statistical language learning account and the interactive, or constraint-based, models of sentence processing. On the other hand, the conclusions drawn by Reber, Marcus and colleagues, Peña et al. (2002), and Bonatti et al. (2005) all indicate an insignificant role of statistics in acquisition, and therefore implicitly support the idea that the only information used in (initial stage) sentence comprehension is abstract syntactic categories-since the system seldom, if ever, make use of statistical information in acquisition, there is no resulting statistical information available in the system during comprehension. The latter is

correct, a positive answer is given to the question: "*Does grammar start where statistics stop*?"

The parallel statistical and innate predisposition interpretations of the results speaks to the difficulty of using behavioral data obtained from end-point measures to infer the processes involved in learning and use and the need for a more sensitive measure. ERP provides such as approach because of its capacity to monitor performance on-line and its high temporal resolution. This methodological advantage is accompanied by a better understanding of the properties of some ERP components (such as P600) that can be elicited under certain language conditions. The discussion in the next section focuses on ERP research relevant to the present study.

2.3 Event-related Potential (ERP) Studies on P600

In this section, we provided a review of previous studies on the P600, an ERP component that is often referred to as syntactic positive shift (SPS). First we focused on conventional views of the indexing function and antecedent conditions of the P600 component from language studies and contrast this body of literature with Patel and colleagues' work on musical perception and ERP studies on second language acquisition. In doing so we showed the results from their work could critically influence our understanding of the P600 component. However, the significance of this work has largely been ignored in language studies. Specifically, a conclusion we would like to draw from these studies concerns three characteristics of the P600: it is (a) experience-based; (b) not specific to language, and (c) its magnitude varies with the degree of deviation from expectation. The intriguing concordance between the characteristics of this syntactic-associated ERP component and the core features of the statistical language learning discussed in section 2.1 provides evidence favoring the statistical learning account of grammar. The work by Patel and colleagues is valuable in re-thinking and re-evaluating earlier language studies on the P600. Specifically, we focused on a previous language

study by Osterhout, Holcomb, and Swinney (1994). In the end, we laid out some concerns and limitations in both Osterhout et al. (1994) and Patel and colleagues' work that needs to be further addressed as well how the P600 component can be used to examine the core questions asked in this thesis.

One of the most intensively studied ERP components in electrophysiological research on language comprehension is the P600 component- a syntax-related ERP component that has been reported to be elicited by antecedent conditions involving a wide range of syntactic violations, including morphological violations of number, gender, and case (Atchley et al., 2006; Coulson et al., 1998; Osterhout and Mobley, 1995), phrase structure (Hagoort et al., 1993; Neville et al., 1991), subjacency (McKinnon & Osterhout 1996; Neville et al., 1991), and subcategorization (Ainsworth-Darnell et al., 1998; Osterhout & Holcomb, 1992). The P600 usually starts at about 500 msec after the onset of the violation, peaks at around 600 msec, and lasts for about 500 msec (see Figure B1). In addition, it is largest over centroparietal scalp region. Based on the polarity and the latency of its maximal amplitude this effect was originally referred to as the P600 (Osterhout and Holcomb, 1992). The P600 is also referred to as Syntactic Positive Shift (SPS), a term that stems from its unique responses to syntactic violation. As argued by Osterhout, McKinnon, Bersick and Corey (1996), the qualitative differences between the P600 and brain responses to semantic and pragmatic anomalies suggested that this is a domain specific response for a modularly organized language processor for syntactic processing.

Early work on the P600 component was conducted exclusively within the language (sentence-level) context. More recently, researchers have reported P600 effects in the area of musical perception (Besson & Faita, 1995; Koelsch, Gunter, Wittfoth, & Sammler, 2005; Janata, 1995; Patel, Gibson, Ratner, Besson, & Holcomb, 1998; Patel, 1998; Steinbeis & Koelsch, 2007). The findings of P600 effects elicited by non-linguistic materials immediately called for re-evaluation of the domain-specific claim of this

syntactic-related component. The most striking evidence came from a study by Patel and colleagues (1998). They compared ERPs elicited by structural incongruities in language and music in a group of adult participants. Participants in their study were presented with musical phrases from 3 different conditions varying in chord sequences: an in-key chord condition, a nearby-key chord condition, and a distant-key chord condition. As a result, the harmonic incongruity for target chords in a phrase was systematically manipulated such that the most congruous (or harmonic) target chords were always the in-key chords, then the nearby-key chords, and finally the distant-key chords. They found a significant hierarchy of the P600 effects: the P600 elicited by the nearby-key chord condition was significantly more positive than that elicited by the in-key chord condition; and the P600 elicited by the distant-key chord condition was significantly more positive than that elicited by the nearby-key chord condition. When the same participants were presented with English sentences containing syntactic incongruities, a hierarchy of the P600 effects was also found among sentences of three different sentence conditions: the largest P600 effects were found in the ungrammatical sentence condition (e.g., *Some of the senators endorsed the promoted an old idea of justice*), followed by an ambiguous sentence condition (e.g., *Some of the senators endorsed promoted an old idea of justice*), which in turn was followed by a well-formed simple unambiguous sentence condition (e.g., *Some of the senators had promoted an old idea of justice*). Critically, the P600 effects elicited by structural incongruities in the language and music domains were statistically indistinguishable, both in terms of amplitude and scalp distribution in the P600 latency range. Patel et al. concluded that the indistinguishable effects of the P600 elicited by linguistic and musical materials from the same participants revealed that the P600 is not the signature of a uniquely linguistic process. Instead, the finding suggested a strong overlap of cerebral structures and neural processes involved in the processing of musical and linguistic structures (Patel, 2003).

In addition to their contribution to the domain-independence of the P600, it should be noted that the participants in Patel et al. (1998) study were those with extensive prior musical experience, which implies a role for experience in the P600 responses. There is no doubt that their participants also had extensive prior experience with their native language. In fact, there is evidence both in the field of musical perception and second language acquisition pointing to an experience-based characteristic of the P600. Besson and Faita (1995) compared harmonic violation in musical perception and found that the participants' familiarity with the melodies affected the amplitude of a late positive component peaking around 600 msec.

In the Patel et al. study, a hierarchy of the P600 magnitude was observed both in the musical and language tasks. Given that the P600 seems to be experience-based, can it be that the magnitude of the P600 is simply reflective of the amount of experience one has had with a certain kind of stimuli? In other words, could the magnitude variations among the 3 language conditions or the musical conditions in their study be explained because the participants had had the least prior experience with ungrammatical sentences/ distant-key chords and the most experience with well-formed unambiguous sentences/in-key chords? The cognitive processes that give rise to the graded P600 effects were not explicitly stated in the 1998 study. However, later in a discussion article Patel (2003) suggested that the P600 is an index of integration cost in both linguistic and musical syntactic processing, and syntactic integration is more costly when the occurrence of the current element was not *implied* by past elements. In this view, the magnitude variations of the P600 could be interpreted as "goodness of fit" between a target structure and an expected structure projected from past elements. This interpretation is in line with the prior experience interpretation but is more sophisticated in that it involves a process of expectation generation. The essence of expectation generation is an associative relationship between two or more elements that are more or less dependent to each other. This is reminiscent of the statistical language learning account discussed earlier as

statistical representations allow for graded or probabilistically ranked expectations, whereas algebraic rules have considerable difficulty with such graded expectations. For the language task in Patel et al. (1998) study, the magnitude of the P600 is therefore reflective of the differences between activation states of structures projected (or expected) based on past elements and the target activation states based on the target structures. Adjustment of activation levels according to the degree of "*deviation*" of activation states is therefore not inconsistent with the idea of resource "cost" of integration.

The domain-independent, experience-based, and statistically-based expectation generation nature of the P600 found in sentence processing favors the statistical language learning account of grammatical acquisition. This calls for re-thinking and re-evaluating the P600 effects reported in earlier studies focused on language. Although the sentences used to construct the 3 language conditions in Patel et al. (1998) were not specifically selected based on input statistics, one earlier language study on the P600 by Osterhout, Holcomb, and Swinney (1994) provides a sufficiently close experimental setting to further evaluate the proposal that the amplitude of the P600 is associated with the degree of deviation of a target structure from a expected structure based on the probabilities between the two. In Osterhout et al. (1994) study, they examined the P600 effects in adults' sentence reading. While all the experimental sentences had a sentential complement (SC) structure (e.g., *The traveler saw the island was deserted*), the matrix verbs (*saw*) of those sentences varied in verb transitivity, resulting in four principal experimental conditions: intransitive matrix verbs with a SC structure (e.g., *She agreed the actress was great*), intransitively-biased verbs with a SC structure (e.g., *She knew the actress was great*), transitively-biased verbs with a SC structure (e.g., *She forgot the actress was great*), and transitive verbs with a SC structure (e.g., **She bought the actress was great*). All the sentences in the *transitive* condition were ungrammatical. Although all the sentences in the *intransitively-biased* and the *transitively-biased* conditions were

grammatically well-formed sentences, the post-verbal noun phrases in these two biased conditions were temporary ambiguous between a direct object reading (*She knew the actress*; *She forgot the actress*) and a subject noun of a sentential complement (…*the actress was great*). The verbs used in Osterhout et al. study were selected based on verb bias ratings obtained from a sentence completion task published by Connine, Ferreira, Jones, Clifton, and Frazier (1984). Most relevant to the current discussion, Osterhout et al. found that the magnitude of the P600 varied among the four sentence conditions. Similar to the magnitude hierarchy of the P600 reported by Patel et al., the mean amplitude of the P600 for auxiliary verbs (the disambiguated region) in the *transitive* condition was more positive going than that in the *transitively-biased* condition, which was in turn more positive going than that in the *intransitively-biased* condition. Mean amplitude differences between *intransitively-biased* and *intransitive* conditions were not significant. Based on this finding, Osterhout et al. suggested that the parser can use verb subcategorization information to resolve local ambiguity. However, given that the verbs used in their study were selected based on verb bias ratings obtained from a sentence completion task, the data from their study also provided direct evidence for the influence of input statistics on sentence comprehension. This is in accordance with the proposal suggested by Patel and colleagues. Although Osterhout et al. speculated that the P600 amplitude was associated with syntactic expectations engendered by the preceding context (see also Osterhout & Nicol, 1999), they concluded that the amplitude variation of the P600 was a function of the perceived syntactic well-formedness of the sentence. As discussed below, these conflicting proposals probably arose from the results they obtained from a behavioral measure (acceptability judgments).

Based on the studies discussed above, it can be conclude that the syntactic-related ERP component is (a) experience-based; (b) not specific to language, and (c) its magnitude varies with the degree of deviation from expectation. These are in fact characteristics core to the statistical language learning account. The finding of graded

P600 effects is of importance to the current study in that one critical difference between statistical representations and algebra-like rules of grammar is that the former allow for graded or probabilistically ranked expectations, whereas the later has considerable difficulty in accounting for such graded effects since rules are operations of abstract unitary symbols. Still, there are some concerns that need to be further addressed before a more certain conclusion can be drawn.

First of all, in the Patel et al. study the graded magnitude of the P600 was found among 3 conditions for both their musical and language tasks. One potential criticism to the claim of graded P600 based on their results is that the intermediate condition in Patel et al. study (i.e., the nearby-key chord condition in their musical task and the ambiguous sentence condition in their language task) could have been due to participants' confusion with the materials. For instance, for those ambiguous sentences, it is possible that their participants considered the sentences well-formed or grammatical 50% of the time and ill-formed or ungrammatical 50% of the time, depending on whether they were initially garden-pathed and whether they were able to successfully recover from initial incorrect processing. If this mixture of responses is true, then the varied amplitude of the P600 reported by Patel et al. (1998) was not truly graded but instead categorical, and the intermediate magnitude was the averaging of brain responses to both grammatical and ungrammatical conditions. This is problematic since it would suggest symbolic rule-based nature of sentence processing. The use of four different sentence conditions by Osterhout et al. (1994) should have avoided this problem in that the magnitude of the P600 elicited by sentences from the two intermediate conditions (i.e., the two differently biased conditions) was significantly different. The argument is that if the P600 effects elicited by the sentences from the intermediate conditions (i.e., the two biased conditions) were due to a mixture of categorical representations, then no differences should have been obtained between the two differently biased conditions. However, their behavioral data did not seem to directly support this argument. During the experiment, Osterhout et

al. (1994) asked their participants to provide acceptability judgments of the sentences. Acceptable sentences were defined as semantically coherent and grammatically correct. The resulting behavioral data for the *intransitive*, *intransitively-biased*, *transitively-biased*, and the *transitive conditions* were: 91%, 84%, 66%, and 4%. If their judgments are to reflect grammaticality to a significant degree, it would be reasonable to expect higher acceptability rates (at least close to the *transitive* condition) for the two biased conditions. This is because the participants gave acceptability judgments at the *end* of reading each sentence. By the end, temporary uncertainty should have been resolved and the biased sentences should be grammatical, making this an offline measure. Since the participants in Osterhout et al. study were instructed to provide responses based on both semantic and syntactic aspects of the sentences, it is uncertain to what extent the participants' acceptability judgments reflected grammatical aspects of knowledge or semantic knowledge. If the pattern of the behavioral data reflected the participants' grammaticality judgments to a significant degree, then we have to seriously consider whether the intermediate magnitude of the P600 from the two biased conditions reflected a mixture of categorical brain responses to grammatical and "*ungrammatical*" items.

The same concern also exists when inspecting the behavioral data reported in Patel et al. (1998) study. The participants judged well-formed grammatical sentences, ambiguous sentences, and ungrammatical sentences acceptable on 95%, 61%, and 4% of the trials. The definition of acceptability was the same as that in Osterhout et al. study. The behavioral data from the musical task was of the same pattern. Again, it is not certain to what the extent the behavioral data reflected judgments based on the grammatical aspects of the sentences. However, this is critical in order to tease apart statistical and algebraic rule-based representations of grammar.

The second concern pertains to verb categorization in Osterhout et al. (1994) study. The verbs used in Osterhout et al. study were selected based on published norms reported by Connine et al. (1984). In Connine et al. study, adult participants received a

list of verbs and were asked to write a sentence for each verb. The resulting sentences were assigned to a categorizing scheme, in which 6 syntactic frames were further categorized as transitive frames (such as V + NP as in "T*he teacher remembered his books*" ; V+ NP +NP as in "*The teacher have Tom his books*") and the other 6 frames intransitive frames (such as Zero Post-verbal Noun Phrase as in "*The teacher remembered*"; V + PP as in "*The photographer sat in the room*). Verb bias ratings were then calculated for each verb by dividing the frequency of transitive usage by the sum frequency of transitive and intransitive usage. The verbs that Osterhout et al. used to construct their 4 transitivity conditions (*intransitive*, *intransitively-biased*, *transitively-biased*, and transitive) were based on the verb transitivity ratings obtained from the sentence completion task in Connine et al. study.

There are two methodological concerns with this approach. First of all, one of the principal questions under examination in Osterhout et al. study was if lexical subcategorization information contributed to ambiguity resolution. All the experimental sentences in their study were sentences with a sentential complement clause (e.g., *She remembered the actress was great*). With this design, the post-verbal nouns (*actress*) constituted the principal "ambiguous region" they could be temporary interpreted as the direct object (DO) of the matrix verb (*remembered*) or the subject of the sentential complement clause (SC). This type of ambiguity constituted a main source of syntactic structure competition, and the key question asked in their study was if the preference of the matrix verb (*remembered*) exerted influence on this competition. Given the goal of their study and the purpose of the current study, it is more appropriate to determine the influence of the matrix verbs on subsequent structural competition by only considering possible subsequent (structure) candidates. For instance, in hearing or reading a sentence like "*She remembered the actress was great*", the transitivity information of the matrix verb (*knew*) should only affect the competition between DO/SC reading when there is a post-verbal noun phrase. Compare this example with another sentence "*She remembered*

to sign the paper". However, the sentence involves the same matrix verb (or the same expectation generator) at the point when the infinitival (*to*) was unfolded during reading or listening, DO/SC competition no longer exists. In addition, according to the categorization reported in the study by Connine and colleagues, both of the above sentences were members of the intransitive category, although sentences of this type do not have a post-verbal noun phrase and therefore do not directly contribute to DO/SC competition. Therefore, if we are to examine how the statistics of the matrix verb and the expectations generated based on that verb's statistics influence subsequent processing of sentences with a post-verbal noun phrase, then it is more appropriate for the transitivity status of a verb to be based on its probabilities of occurrence in DO/SC constructions.

Second, the selection of verbs based on data from a sentence completion task would potentially inflate the frequency of the most probable usage of a given verb. In the study by Connine and colleagues, the verb transitivity rating for each verb was based on only one response from each participant. For verbs that can be used in multiple constructions, it is conceivable that the construction with the highest probability in the array of possible constructions was more likely to be used to form a sentence. As a result, the DO/SC probabilities or the transitive/Intransitive probabilities elicited from such a task might not accurately reflect the probabilities of different constructions for the verbs. This is especially critical for the biased verbs since these verbs were those which could occur in both DO and SC constructions. More importantly, the mean amplitude differences of the P600 between these two conditions is key to evaluate if the graded effects of the P600 were due to statistical representations or a mixture of categorical responses (grammatical/well-formed versus ungrammatical/ill-formed) that could have resulted from operation of unitary symbolic rules.

2.4 Summary and the Current Study

In section 2.2, we summarized issues central to the on-going debate between different perspectives on language acquisition and the implications these different perspectives have on comprehension. One difficulty in teasing apart the two theoretical families has to do with the use of behavioral data to reason back to the underlying processing that give rise to the observed performance. In section 2.3, we reviewed previous ERP studies on sentence processing with a particular focus on processing ambiguous sentences, i.e., sentences containing points of uncertainty and therefore providing a situation for differential expectancy generation. Early studies focusing on the P600 or so-called Syntactic Positive Shift (SPS) were largely restricted to language studies. More recently, studies have provided evidence showing that violation of harmonic expectation in music could also elicit a P600 that is statistically indistinguishable in terms of latency range and topography from the P600 elicited by syntactic incongruity in processing sentences. Together with the finding that the amplitude of the P600 seems to vary consistently with the distance between the expected and the target outcomes, this non-domain specific, graded nature of sentence processing fits well with the statistical language learning account.

In the current study, ERP measures in a natural language sentence processing task and an AGL were conducted to examine the questions. The rationale of examining sentence processing in adult in the sentence processing task is that if statistical learning is not at work in grammatical acquisition, then sentence processing in mature systems should not exhibit any sign of statistics. The opposite would hold for statistical language learning. Additionally, an AGL task was also included. This allowed a more direct inspection of the influence of input statistics on later processing. Comparisons of the scalp distribution from the two tasks would provide a piece of information to the debate of statistical versus algebra or rule-governed nature of learning in a typical artificial language learning task as discussed in section 2.2.

In the sentence processing task, we replicated the study by Osterhout et al. (1994). In the original study, the categorization of verbs to different transitivity subcategories was based on the frequency of each of these verbs occurring in transitive versus intransitive syntactic frames in a sentence completion task conducted by Connine et al. (1984). However, since not all (transitive or intransitive) syntactic frames are considered during expectancy generation in the (ambiguous) region that prompts potential competition of multiple candidates, it is thus more reasonable to calculate *direct object* versus *sentential complement* probabilities (DO/SC probabilities) rather than transitive versus intransitive probabilities. Given the reason plus the frequency counts for some of the verbs selected by Osterhout et al. were missing in Connine et al.'s report, a multi-corpus analysis of the SC/DO probabilities of the 20 verbs used in their study was conducted first. These verbs were then assigned to one of four groups (*intransitive*, *intransitively-biased*, *transitively-biased*, and *transitive*) according to their transitivity bias based on the results from the corpus analysis. A set of sentences containing the verbs with varied DO/SC probabilities but were all in a SC construction constituted the experimental sentences. Therefore, the goodness of fit is the highest for the verbs in the *intransitive* category, then the verbs the *intransitively-biased* category, then the verbs in the *transitively-biased* category, and finally the verbs in the *transitive category*.

As discussed earlier, grammar in the generative/minimalist paradigm is considered as idealization that abstracts away from a variety of performance factors, and the resulting abstract rules allow people to distinguish *grammatical* from *ungrammatical* sentences. Given that P600 is sensitive to syntactic incongruity, this account would predict categorical P600 patterns: significant P600 responses elicited by sentences containing a transitive matrix verb (*ungrammatical* sentences) versus a lack of significant P600 responses for the other three conditions (*grammatical* sentences). On the other hand, if the statistical language learning account is correct, then graded P600 (mean amplitude) ranked according to the distance between the expectancy generating from the

verbs and the resulting construction (i.e., SC) would be expected, with the sentences in the transitive category elicited the largest P600, then the transitively-biased category, and then the intransitive-based category and the intransitive category. Parallel predictions also apply to the corresponding sentence processing models compatible with the two contrastive accounts of language acquisition.

The AGL task was designed in a way such that it has similar manipulations to test categorical versus graded ERPs as the sentence processing task. The natural sentence processing task taps acquired knowledge, whereas the statistical language learning task allows performance to be monitored overtime and therefore differences in performance observed in a post-training phase as compared to performance in a pre-training (baseline) are unlikely due to knowledge (either innate or based on prior language experience) that the participants bring to the task. In addition, topographic analyses were conducted to examine if there are topographic differences in P00 amplitude that would suggest that processing sentences and processing artificial language were subserved by different neural structures.

CHAPTER III

ARTIFICAL GRAMMAR LEARNING TASK AND NATURAL LANGUAGE SENETNCE PROCESSING TASK

In chapter 2, we summarized previous work on statistical learning of grammar and argued that this domain-general *language learning* account provides a coherent theoretical ground for on-line sentence processing (*language use*). Further support for the role of statistical learning in grammatical acquisition was drawn from review of neurophysiological studies on natural language sentence processing. Focusing on the SPS/P600 component, we argued that this ERP component is characterized by what are considered fundamental features of statistical learning. This includes the domain generality of the P600 for the elicitation conditions are not specific to language materials, its association with experience for the magnitude of the P600 components elicited from experienced participants of a given type of stimuli differs from that from less experienced ones, and its sensitivity to input statistics for the magnitude of the P600 component varies systematically with input probabilities.

The finding that the magnitude of the P600 component varies systematically with the likelihood of a target element and already presented ones (or goodness of fit) opens a window to directly examine if magnitude variations of P600 responses are given rise to learning input statistics in the first place. To this end, an artificial grammar learning (AGL) task was used in the current study. We compared ERPs obtained from the same group of participants before training and after they showed high degree of knowledge of this artificial language. This comparison would provide information to ***whether the P600 effect associates with the experience or the state of knowledge of learners***? In addition,

the training sentences contained materials varying in conditional probabilities between a target word and already unfolded words. This allows for manipulation of the *goodness of fit* of sentences used in the post-training test. In the test phase, we used sentences of 4 conditions, Excellent Fit (p = 1.0), Good Fit (p = .67), Low Fit (p = .33), and Bad Fit (p = 0) to obtained ERP data. Comparisons of ERP data across conditions would provide information to question that directly addresses the primary goal of this thesis: ***does the magnitude of the P600 vary systematically with the conditional probability of the input after learning?***

Even we are able to obtain P600 effects that are associated with to the learners' experience to the artificial language and show varied magnitudes of the P600 effects due to input statistics, statistical learning of artificial grammar does not necessarily justify the argument of the same kind of learning involved in natural language acquisition. Our solution is to seek evidence from a natural language sentence processing (NLSP) task. We replicated the same task used in Osterhout et al. (1994) study (Experiment 2). Similar to the AGL task in which test sentences varying in the goodness of fit, we collected ERP data elicited by natural language sentences of varied conditional probabilities. We compared the ERP data obtained in the NLSP and the AGL task to examine the third question asked in this thesis: ***does the P600 obtained in the AGL task conform to the P600 obtained in the natural language task?*** Specifically, we examined if the same pattern of magnitude variations of P600 effects seen in the AGL task also show in the NLSP task. Furthermore, we compared topographic distributions of P600 effects in the two tasks to examine if the same neural mechanisms were recruited for syntactic processing of natural language sentences and statistical learning of sequential patterns.

In section 3.1, we provided detailed information about the AGL task and sentences used in each of the pre-training, training, and post-training session. In section 3.2, we summarized the original task used in Osterhout et al. (1994) study discussed three major modifications of their task to suite the purpose of the current study. Results from a

corpus analysis were also included in 3.2. In this thesis, each participant took part in the NLSP and the AGL tasks. The use of a repeated measure design is to provide a more common basis for comparisons of the two tasks. Section 3.3 included a detailed description of the overall procedures of the whole study. In section 3.4, predictions of the two experiments were discussed. Peña and colleagues argued that learning grammatical structures relies entirely on algebraic computations for statistical learning is incapable of learning abstract linguistic structures, which echoes Marcus and colleagues' proposal of symbolic rule-based grammatical learning and the traditional generative/minimalist paradigm on grammatical acquisition. This is the general theoretical stance that contrasts with the statistical learning account. Therefore, the predictions laid out in section 3.4 were organized with the purpose to show the respective predictions of the two contrastive theoretical accounts for grammatical acquisition.

3.1 Artificial Grammar Learning (AGL)

In this section, the AGL task used in this thesis was discussed. We first illustrated the overall structure of the artificial grammar, and then described in more details features of this artificial grammar that are critical for testing the questions of interest. Sentence strings used in different phases, including a pre-training, training, and post-training phases, and overall procedures of this task were also discussed.

3.1.1 Overall Structure of the Artificial Language

A state diagram of the miniature grammar and the corresponding tree diagram of the miniature grammar used in the present study are displayed in Figure B2. This grammar captures the relationships between nodes by providing separate, alternative paths through the network. Every sentence starts from the beginning node ([) to the end node (]). Single letters within the nodes represent word classes and the arrows indicate valid transitions between nodes. Every sentence must start with one D word and end with an E, L, or N word. In addition, each sentence must include either a C word or an E word.

An A word is optional, but is always preceded by a D word and followed by an N word. The language can generate 12 different possible sentence types (e.g., DANE, DNEL), ranging in length from 3 to 8 words. Only sentences of six or fewer words, of which there were 8 different sentence types, were used in the experiment (for either training or test). In total, the language can produce 1,888 possible sentences (a total of 608 sentences of 6 or fewer words).

The vocabulary of this language consisted of 15 nonsense words (see also Figure B2). Two words each were assigned to D and A, four words to N, and three words to L. For C and E, each of which has a prototypical (C_1, E_1) and a peripheral (C_2, E_2) word members.

Critical to the current study were the distributional differences among C_1, E_1, C_2, and E_2. First, let's focus on the two prototypical word tokens C_1 and E_1 .These two prototypical word tokens always follow the paths specified in Figure B2: the presence of C_1 always predicts a subsequent D word (*probability = 1.0*). No such relationship holds between E_1 and D words (*probability =0*). Different from their prototypical word member, the two peripheral word tokens have more varied statistical patterns. To more clearly illustrate this difference, Figure B3 demonstrates the statistical relationship between C_2 and valid subsequent paths. For the peripheral word token C_2, it flows through the diagram following the same routes valid for C_1, the prototypical member of the C class, 67% of the time. The rest 33% of the time C_2 follows the routes valid for E_1, the prototypical member of the E class. The reverse probabilistic patterns held true for E_2, as shown in Figure B4. E_2 takes the routes valid for E_1 67% of the time and the routes valid for C_1 33% of the time.

These varied transitional probabilities gave rise to probability differences among the 4 lexical items in class C and class E. These manipulations are analogical to the distribution differences of verbs of different transitivity in natural languages. For instance, E_1 is never followed by D(A)N in this grammar and therefore its distribution is

similar to a highly intransitive verb. The opposite is true for C_1 as its appearance always predicts a subsequent D(A)N(L) phrase. For the two peripheral cases C_2, and E_2, although they could occur in structures valid for C1 and E1, their distributions biased toward the class's distributions and therefore they act in a way similar to transitively- or intransitively-biased verbs in natural language. It is important to note during learning of the AGL, we did not provide our participants any visual reference to match sentence strings. Nor did we provide our participants any information about word categories. Therefore, it is unlikely that our participants would assign a specific category such as VERB, ADJECTIVE or NOUN to the C or the E class. The use of "*verb transitivity*" here is to provide an example to illustrate why such probabilistic differences are sensible manipulations under the context of studying language learning.

During training, the participants listened to and repeated sentences containing one of the 4 keywords following their own conditional probability structures. In the post-training phase, the major test sentences involved all 4 words occurred in structures valid for C_1. This led to graded degrees of variations in terms of *goodness of fit* of the 4 key words and the syntactic structures in which they occurred: E1 had the worst goodness of fit with the structures (the ***Bad Fit*** condition), which is followed by E2 (the ***Low Fit*** condition), which is in turn followed by C2 (the ***Good Fit*** condition). The goodness of fit between the key word and the structures is the best for C1 (the ***Excellent Fit*** condition).

3.1.2 AGL Training Sentences

Sentences used in training were 192 sentence**s** randomly selected from a sentence pool consisting of 608 sentences of 6 or fewer words. Each participant received up to 4 training sessions. At the end of each training session, participants provided grammaticality judgments for a set of 24 questions. A criterion of 90% accuracy was used to determine if the participants succeeded or failed the AGL task.

Each training session consisted of four phases. In the first phase, participants were familiarized with the vocabulary of this language by first listening to the 15 novel words three times while the novel words were displayed on a computer screen. This was followed by a vocabulary repetition session and another vocabulary listening session. In the second phase, participants first listened to a set of short phrases of this language. This was followed by repetition of the same set of phrases for two times. Structures included in generating the phrases for this training sub-phase were D-N, D-A-N, D-N-L, and D-A-N-L. In the third phase, participants were presented with longer phrases, including phrases in the structures of C/E-D-N, C/E-D-A-N, C/E-D-N-L, and C/E-D-A-N-L. Probabilities of occurrence with subsequent phrases were controlled for individual C words and E words such that they followed the pre-determined conditional probabilities in the grammar. In the fourth phase, full sentences were presented to participants. Sentences presented in this phase consisted of 48 sentences for each C word and each E words (a total of 192 sentences).

At the end of each training session, participants provided grammaticality judgments for 24 sentences. Their performance was then evaluated with a criterion of 90% accuracy to determine if a participant required more training. The same training materials were repeated in each training visit for up to 4 training visits.

3.1.3 AGL Pre- and Post-training Test Strings

An additional 384 experimental (grammatical) sentences, 96 sentences for each word in class C and E, were selected from the sentence pool to form stimuli used in the post-training test. None of the sentences used in the test session occurred in the training.

The test sentences were all in structures valid for the prototypical word token C_1. Therefore, the probability of occurrence of a key word in these structures was 100% for C_1 (the Excellent fit condition), 66% for C_2 (the Good fit condition), 33% for E_2 (the Low fit condition), and 0% for E_1 (the Bad fit condition). The ERP measure was time-locked

to the next word of each of the 4 key words. To provide a good baseline for the ERP data analyses, the sentences for comparisons were kept acoustically identical up to the point where violation appeared. At the end of each sentence, participants also provided grammaticality judgments for each sentence.

In addition to these test sentences, a set of 384 filler sentences were also included in the post-training test. There were 96 filler sentences with C_1 in structures grammatical for the prototypical word token E_1 (ungrammatical), 96 filler sentences with E_1 in structures grammatical for the prototypical word token E_1 (grammatical), 96 filler sentences with C_2 in structures grammatical for the prototypical word token E_1 (grammatical, less preferable), and 96 filler sentences with E_2 in structures grammatical for the prototypical word token E_1 (grammatical, preferable). The purpose of including a set of filler sentences of the same amount as the test sentences but in different syntactic structures were to prevent context-specific prediction effects due to repeated use of the same types of sentence structures.

An additional 384 ungrammatical sentences caused by a scrambled word order or a missing word were also included such that the test questions had a balanced number of grammatical (50%) versus ungrammatical (50%) sentences. The whole set of stimuli used in the (pre- and post-) test phase consisted of a total of 1152 sentences, which were randomly ordered and then divided into 12 blocks of presentation.

The same set of sentences used in post-training test was also used to obtain EPRs and grammaticality judgments before training started. Previous studies on statistical learning have demonstrated that the phonological properties of the stimuli could exert an influence on learning and therefore would confound with statistical learning of the stimuli (Onnis, Monaghan, & Chater, 200; Peña, Bonatti, Nespor, Mehler, 2002). To deal with this concern, the participants in the current study also gave grammaticality judgments to the same set of test stimuli used in the post-training test before they received any training on the artificial grammar.

3.1.4 Procedure

3.1.4.1 Pre-training Test

To assure that the expected ERP differences were actually caused by differences in syntactic processing, a pre-training test session in which participants provided grammaticality judgments for the same set of sentences used in the post-training test was given before training starts. The EEG was recorded during this pre-training phase.

During the pre-training test, participants were asked to provide grammaticality judgments for 12 blocks of sentence strings (96 sentences for each block) generated by the artificial language. Participants were instructed to give their responses based on their intuition by pressing a Yes (*grammatical*) or No (*ungrammatical*) button on a computer keyboard. Participants were seated comfortably in a quiet room approximately 4 feet from a computer screen. The participant's left and right index fingers were each positioned over the left and right buttons, one corresponded to a Yes (*grammatical*) response and the other corresponded to a No (*ungrammatical*) response, on a computer keyboard held in the lap. Six of the participants whose data were included in the final data analyses had Yes button on the left and the other six had the No button on the left. Before the session started, participants were visually presented with a display of the EEG on which they observed the effects of blinking, sniffing, jaw movement, and eye movements, and were given specific instructions to limit such behavioral throughout the experiments.

The sentences used in the pretest phase were the same sets of sentences used in the final test phase. Each trial consisted of the following events: 1) A fixation cross appeared in the center of the screen 150 msec before the onset of the first word; 2) A sentence was presented auditorily, one word at a time with a 50 msec between-word pause, with the fixation remained on the screen for 1000 msec after the onset of the last word. Participants were asked not to move their eyes or blink while the fixation cross was

present; 3) After listening to each sentence, a Yes/No prompt appeared on the screen asking the participants to decide if the sentence was a “grammatical” or “ungrammatical” sentence. The locations (left or right) of the Yes/No on the screen corresponded to the locations of the Yes/No buttons on the keyboard. Participants were instructed to give their response as soon as possible once the Yes/No prompt appeared on the screen; 4) Following the Yes/No response, a black screen appeared and the participants were told that they can blink or take a short break before they pressed the space bar to listen to the next sentence. The pre-training test session lasted approximately 2.5 hrs.

3.1.4.2 Training

After the pretest, participants came back on a different day for the first training session. Each participant was given up to 4 training visits to reach a criterion of 90% of accuracy in judging the grammaticality of sentences generated with the artificial grammar. The first training session was conducted within a week from the pretest session, and each training visit was conducted on consecutive days or 2 days apart at most. The participants were informed that their goal was to learn to differentiate correct sentences from incorrect session to a degree of 90% of accuracy within 4 learning visits, each of which lasted for approximately an hour. To encourage learning, participants who passed the criterion in less than 4 visits were provided with the same amount of compensation as 4 visits. Once the criterion was reached, participants proceeded to a final Test phase in which they provided grammaticality judgment to the same 12 blocks of sentences used in the pretest phase. The final test phase was conducted within a time frame of 3 days from the last training visit. Participants who failed to reach 90% of accuracy on the fourth learning visit discontinued participation.

3.1.4.3 Post-training Test

Overall, the final test phase followed the same procedures as the pretest phase, except for that on the final test session the participants were only verbally reminded with

the adverse effects of artifacts due to blinking, sniffing, jaw movement, and eye movements.

3.1.5 Participants

23 participants (11 male, age range 20-62) were recruited from the University of Iowa campus. All these participants were right-handed native speakers of English and none of whom have a history of language, hearing, or neurological impairment.

3.1.6 ERP Data Acquisition

The EEG was recorded from 36 electrodes using Ag-AgCl electrodes attached to an elastic cap (see Figure B5). Thirty electrodes were positioned over homologous location of the two hemispheres. Location were as follows: lateral sites F7/F8, FT7/FT8, T7/T8, TP7/TP8, P7/P8; midlateral sites FP1/FP2, F3/F4, FC3/FC4,C3/C4, CP3/CP4, P3/P4, O1/O2; and midline sites FZ, FCZ, CZ, CPZ, PZ, OZ. The channels were referenced to an electrode placed on the left and right mastoid. Horizon eye movement was monitored via electrodes placed over the left and right outer canthi. Vertical eye movements were monitored with electrodes over the left inferior and superior orbital ridge. All electrode impedances were kept to 5 kohms or fewer. The electrical signals were amplified with a bandpass of 0.1 and 100-Hz and were digitized online (Neuroscan 4.4) at a rate of 500 Hz.

The continuous EEG was segmented into epochs in the interval of -200 msec to +1000 msec with respect to the onset of the target word. ERPs were time-locked to the onset of the target word, i.e., the next word following C1, C2, E1, or E2. Trials with eye-movement artifacts were excluded from the average. ERPs were baseline-corrected with respect to the 200-msec pre-stimulus interval and referenced to an average reference. Separate ERPs were computed for each subject, each condition, and each electrode.

3.2 Natural Language Sentence Processing (NLSP) Task

This section is composed of two parts. In section 3.2.1, we will summarize the original sentence processing task used in Osterhout et al. study (Experiment 2) and discuss three major modifications we made to suit the purpose of this thesis. In section 3.2.2, we will report results from a corpus analysis of the 20 verbs used in Osterhout et al. study. Different from Osterhout et al. study in which these verbs were categorized in terms of verb transitivity, we calculated the probabilities of each verb used as a matrix verb in sentences containing a sentential complement clause (***SC probability***) against the probabilities of each verb followed by a direct object (***DO probability***). The resulting SC/DO probabilities served to re-categorize the verbs for later data analyses.

3.2.1 NLSP task

Osterhout et al. examined the P600 effects in adults' sentence reading. The task used in their study (Experiment 2) consisted of 480 target sentences, 120 for each one of the following 4 verb types: transitive verbs, intransitively-biased verbs, transitively-biased verbs, and transitive verbs. Each verb type contained 5 verb members. The intransitive verbs were "agree", "hope", "think", "insist", "decide"; intransitively-biased verbs were "believe"; "know", "promise", "remember", "guess"; transitively-biased verbs were "hear", "forget", "understand", "see", "charge"; and transitive verbs were "buy", "discuss", "follow", "include", "force". The entire set of the experimental sentences from their study is provided in Appendix C.

All of the 480 target sentences were sentences with a sentential complement (SC) clause. In a sentence with a sentential complement clause, such as "*The captain believed the crew was unhappy*", the post-verbal noun phrase (*the crew*) acts as the subject of the SC clause, rather than the object of the matrix verb (*believed*). Therefore, using transitive verbs in such construction would result in an ungrammatical sentence in that transitive verbs require a post-verbal noun phrase as a direct object. Given the 4 types of verbs used

in the task, the resulting 4 conditions are an ***Excellent Fit*** condition, in which intransitive verbs occurred in sentences with a SC clause, a ***Good Fit*** condition, in which intransitively-biased verbs occurred in sentences with a SC clause, a ***Low Fit*** condition, in which transitively-biased verbs occurred in sentences with a SC clause, and a ***Bad Fit*** condition, in which transitive verbs occurred in sentences with a SC clause.

Three modifications of the original task were made to form the version of task used in this thesis. First of all, the original version of the task included a set of 120 filler sentences, in addition to the 480 target sentences. These filler sentences were either sentences of a transitive structure or sentences of a scrambled word order to. The purpose of including a set of filler sentences was to balance the number of Yes/No responses and, probably more importantly, to prevent the structure of the target sentences being predictable due to hearing sentences of the same SC clause all the time. However, the number of the target sentences in their study was three times as great as the amount of the filler sentences. This might account for why in Osterhout et al. study the difference between the Excellent Fit and the Good Fit conditions did not reach significance. During the experiment, their participants heard more sentences with a SC clause. While this particular structure became a predominant one, this material bias could have neutralized the differences in statistical distributions between the Excellent Fit and the Good Fit conditions. Particularly, the more frequent SC structure could have biased their participants toward interpreting the post-verbal noun phrase as a *subject noun* for sentences containing a (intransitively-biased) verb in the Good Fit condition. Since post-verbal noun phrases in the Excellent Fit condition always acted as a subject noun of the SC clause, the task-specific structural bias could therefore diminished ERP differences between the two conditions. In the current version of the task, the number of target and filler sentences was balanced such that 50% of the time our participants heard a sentence in a SC structure and the other 50% of the time in a transitive structure. This control over

task-specific context could therefore eliminate potential influence of task structures and leave performance variations to effects from prior linguistic experience.

Second, instead of using acceptability judgments, we asked our participants to provide grammaticality judgments to deal with a potential problem of "*mixture responses*". In Osterhout et al. study, their participants were asked to judge a sentence acceptable when it was *both* grammatically correct and semantically coherent. As discussed in chapter 2, although acceptability judgments is more naturalistic in the sense that judgments are based on overall acceptability (both semantic and syntactic) of a sentences, the resulting data are often hard to interpret especially if we are to tease apart relative influence of syntax and semantics on the performance. In Osterhout et al. study, although acceptability judgments in the 4 experimental conditions were not statistically evaluated, differences among conditions were observed: 91%, 84%, 66%, and 4% of the target sentences in the Excellent Fit (intransitive), Good Fit (intransitively-biased), Low Fit (transitively-biased), and Bad Fit (transitive) conditions were judged to be acceptable. Since the extent to which syntactic judgments contributed to "unacceptable" responses was unknown, plus examination of ERPs almost always involves some sort of averaging procedure to minimize EEG noise, the graded P600 found in their study might be due to mixture responses, rather than truly reflect an influence of prior language experience.

Taking the Low Fit (transitively-biased) condition for example, assume that syntactic anomaly contributed entirely to the "unacceptable" responses given by their participants, in this case 66% of the test sentences were considered grammatical, while the rest 34% ungrammatical. Therefore, 34% of the sentences in this condition contributed to a P600 effect while the other 66% did not. In this case a significant P600 effect in this condition is a mixture of two types of responses. Compare this with the Bad Fit (transitive) condition in which 96% of trails were considered unacceptable. If the unacceptable responses were entirely due to syntactic anomaly, then when ERPs were obtained by averaging across trials the mean amplitude of a P600 effect would be larger

in the Bad Fit condition than the Low Fit condition since the former has more trials contributing to the P600 effect. This mixture response issue is a critical problem for the magnitude variations of the P600 component constitutes the major dependent variable in this thesis. To cope with this issue, we shifted to a grammaticality judgment task. The use of grammaticality judgments made it possible to filter out ERP trails based on behavioral responses and therefore eliminate the possibility of mixture responses as a source for graded P600 effects.

Finally, different from the Osterhout et al. study in which the sentences were presented in a word-by-word manner on a computer screen, the version of the task used in this thesis contained auditory sentence stimuli. Words that were used to form the target and filler sentences were produced individually such that the same sound token for the post-verbal noun phrase for verbs of the 4 types, eliminating the possibility of intonation differences that might provide a hint on the subsequent structure. To provide a good baseline for the ERP analyses, the four verbs repetitively used in the sentential complete clauses ("*was*", "*would*", "*had*", "*were*") that were time-locked to have the same duration of 566.6 msec.

3.2.2 Corpus Analyses

In this section, we will report results from corpus analyses of the 20 verbs used in Osterhout et al. study. In their study, these verbs were categorized based on verb bias (transitivity) ratings reported by Connine and colleagues (1984). In two sentence completion tasks, Connine and colleagues provided their adult participants a list of verbs and asked them to write a sentence for each verb. Transitivity ratings were then calculated by computing the relative proportions of transitive and intransitive usage. As discussed in chapter 2, verb bias ratings based on only one response from each participant might reflect more of individual differences on the most preferable structure for a given verb than the probabilities a verb used in a given structure, although these two should be

highly correlated. Therefore, we conducted a corpus analysis on the 20 verbs to verify the verb categorizations used in Osterhout et al. study.

All sentences containing these verbs were extracted from two written and one conversational corpora: the Wall Street Journal (WSJ), Brown Corpus (BC), and Switchboard (SWBD). The three corpora vary in size and genre: WSJ is a 1-million word corpus of Dow Jones Newswire stories. BC is the same size, with content from a number of written corpora. SWBD is a spoken corpus (Godfrey, Holliman, & McDaniel, 1992) consisting 1.4 million words. The three corpora were parsed as part of the Penn Treebank Project (Marcus, Santorini, & Marcinkiewicz, 1993). These data are available from the Linguistic Data Consortium at the University of Pennsylvania.

The verbs were classified according to a set of subcategorization frames expanded from the set used by Hare, McRae, and Elman (2003). These constructions were then collapsed into the more general categories of DO, SC and Other according to the summary given in Table A1.

Several criteria were applied during the analyses. First, the frequency of a verb occurring in a particular subcategorization frame was a summed result from frequency calculation of verb root (such as *hope*) and variant of the same root (*hoped*, *hoping*), except for adjectival participles (e.g., the *agreed* price). Second, only completed sentences were included in the analyses. As can be expected, the amount of truncated sentences due to speaker-initiated revision or interruption by conversational partners was relatively higher in the conversational corpus than the other two text-based corpora. The rate of sentence exclusion from the SWBD was 4.65% for the twenty verbs together over the total number of extracted sentences, and was less than 0.1% for the BC and WSJ. Third, passive constructions did not add to the DO count. Although in such construction the “patient” always occurs in the canonical subject position in the sentence, there is no overt post-verbal NP that may be ambiguously treated as DO or SC subject. For the same reason, a post-verbal modifier that eliminated potential ambiguity of the following NP to

be a DO or SC subject were excluded from the SC count. For instance, in the sentence "We agreed *up there* the attitude was unacceptable", the post-verbal phrase "*up there*" eliminated the possibility of the subsequent noun phrase (*the attitude*) being an object of the verb and therefore was exclude from the SC count. On the other hand, sentences like "*We can see only two people are following the regulations*" were included in the SC count since the occurrence of the post-verbal modifier (*only*) did not resolve the ambiguous status of the subsequent noun phrase (*two people*). Finally, only finite embedded clauses were counted as SCs. Infinitival complements were counted as other, as were tensed complements headed with a *wh*-complementizer. Considering first the WH-S construction, these are indeed embedded clauses and therefore raise the interesting question of whether they add to expectation for an SC construction at the verb during on-line sentence processing. However, since they never contain an ambiguous post-verbal NP, the post-verbal *wh*- words should not elicit immediate competition between DO and SC constructions, thus these were counted as Other. The question of how infinitival complements should be classified is less straightforward, as is the issue of their influence on the processing system. Particularly, the post-verbal NP has a number of syntactic properties of an embedded subject noun, and it also behaves much like the DO of the main clause. Hare et al. (2003) viewed the post-verbal NPs in such sentence structure as playing both roles in the course of a sentence derivation. Given the structural indeterminacy, and the fact that there were not sufficient examples to affect the bias counts for any one of the twenty verbs, we followed Hare et al. study and left infinitival complements in the Other category.

The total number of tokens of the 20 verbs varied across corpora (27835 in SWBD: 27835; BC: 8130; WSJ: 6333), and the frequencies of individual verbs varied greatly, from only 4 tokens for the verb "insist" in the SWBD to more than 8500 tokens for the verb "think" in the same corpus.

We first compared the results from the current corpus analysis and responses from the sentence completion task reported in Connine et al. (1984) and adopted by Osterhout et al. (1994). The data reported in Connine et al. (1984) included results from two different normative studies, both of which asked participants (college students, n =78) to write a sentence about a provided topic or setting for a list of verbs (91 verbs in study 1 and 36 new verbs in study 2). The resulting sentences were sorted according to a categorizing scheme, within which were 6 transitive completions and 6 intransitive completions. Transitivity bias was then computed for each verb by dividing the number transitive completions by the number of transitive completions plus intransitive completions. The intransitively-based verbs selected by Osterhout et al. (1994) were those used with clausal complements on 66% of the responses in the sentence completion task, and transitively-biased verbs were those used with a direct object noun phrase on 68% of the responses. There are two methodological differences in verb bias count between the analyses here and Connine et al. study. First of all, transitivity bias in Connine et al. (1984) took into account all intransitive subcategorization frames (including those with or without a post-verbal noun), whereas in our analyses only frames with a post-verbal noun (which are temporarily ambiguous and therefore subject to competition) were considered. The second difference has to do with what subcategorization frames were counted as DO or Intransitive/SC structure. Specifically, in our analyses (Roland & Jurafsky, 1998; Hare et al., 2003) perception complements were included in the DO count, whereas in Connine et al. (and therefore Osterhout et al. (1994) study) perception complements were excluded from the DO count. Additionally, while infinitival complements were excluded from DO/SC count in our analyses, infinitival complements were considered as a type of DO structure in Connine et al. (1984).

To examine the consistency of verb bias across results from the corpus analysis and responses from sentence completion tasks, we first followed the same the DO/SC

categorizations and computing procedures used by Connine et al. (1984). The results are given in Table A2. Theses verbs were organized into 4 categories (intransitive, intransitively-biased, transitively-biased, and transitive) according to Osterhout et al. (1994) study. Overall, the verbs in the intransitive and transitive categories showed expected DO/Intransitive probabilities from the corpus analyses: All five verbs in the intransitive category have an Intransitive probability above 90%. The reversed pattern also applied to the five verbs in the transitive category.

For the two biased categories, the results were less conclusive. For instance, the DO probabilities of “charge” and “see” were perceivably higher than the rest of the verbs in the transitively-biased category and resembled those in the transitive category (mean DO probability was 96.27% and 93.83%, respectively). The percentages of DO were 100% for “charge” and about 96 % for “see” in both the SWBD and WSJ, and relatively lower in the BC (61.54% for “charge” and 85.95% for “see”).

Among verbs originally categorized in the intransitively-biased categories, the DO/Intransitive probabilities were somewhat divergent. For instance, the mean probability of “guess” used in an intransitive frame was 93.86% across the three corpora and above 90% in each of the three corpora, which fell into the same probability range of the verbs in the intransitive category. Conversely, although the verb “remember” showed a bias toward intransitive usage in the 1984 norms, this verb showed a DO probability similar to transitively-biased verbs such as “forget” and “understand” based on the corpus analysis. Further investigation revealed that “remember” was used in a DO structure for more than 50% of the occurrence in each of the three corpora, suggesting that the observed bias toward transitive usage for “remember” did not seem to result from idiosyncratic characteristics of individual corpora.

Given the purpose of the present study, we calculated the probabilities of post-verbal nouns as a direct object (i.e., DO probabilities) against the probabilities of post-verbal nouns as a subject noun (i.e., SC probabilities) for the 20 verbs in each of the three

corpora. Verb transitivity bias was computed for each verb by dividing the number DO or SC by the total number of DO plus SC.

Probabilities of DO/SC structures for the 20 verbs are shown in Table A3. These verbs were organized into 4 categories (intransitive, intransitively-biased, transitively-biased, transitive) according to Osterhout et al. (1994) study. In terms of mean probabilities, all the verbs except for “decide” in the intransitive category showed a SC probability above 90 % if they were followed by a post-verbal noun phrase, and the probability of a subject noun following the verbs in the transitive category was less than 1%. Together, approximately 94 % of post-verbal nouns following the five intransitive verbs were an embedded subject; on the other hand, more than 99% of post-verbal nouns following the five transitive verbs were object nouns in the corpora.

For the two biased categories, the same results were obtained for verbs that showed a different transitivity bias from Connine et al.’s norms in this calculation. This includes, high DO probabilities similar to the verbs in the transitive category for intransitively-biased verbs “charge” and “see”, a high SC probability similar to the verbs in the intransitive category for the intransitively-biased verb “give”, a low SC probability similar to the verbs in the transitive-biased category for the intransitively-biased verb “remember”, and finally a low SC probabilities similar to the verbs in the transitive-biased category for the intransitively-biased verb “promise”.

Given that three corpora together represented language use under various contexts and the size of the sample, we re-categorized verbs according to their DO/SC probabilities from the corpus analysis. This resulted in moving “guess” to the intransitive category and “promise” and “remember” to the transitively-biased category. Three additional verbs-suspect, doubt, and admit-were added into the intransitive-biased category so that each verb category would have the same number of verb exemplars in the new categorization. The three additional verbs were selected from a list of 6 verbs on which we performed the same corpus analyses of their structural preference. The

resulting verb members in each verb transitivity category based on DO/SC probabilities are given in Table A4. This new categorization based on DO/SC probabilities could therefore be described as having a mean of SC probability of 94% for the intransitively verbs (all above 90%), 70 % for the intransitively-biased verbs (between 60-80%) , 25% for the transitively-biased verbs (between 18 to 40%), and between 0.1% for the transitively-biased verbs (all below 10%).

3.2.3 Procedure

Before the task began, the participants were informed that they would listen to 12 blocks of English sentences, and each sentence was extracted from a long story describing events happening in a cartoon world in which everything is animate. This setting was to help the participants to make their grammaticality judgments based on the structure than the semantics of the test stimuli. To further illustrate the cartoon context, examples of grammatically correct but semantically anomalous sentences, such as "*The lady spoon is biting her nails*", and grammatically incorrect but semantically acceptable sentences, such as "*The lady spoon is glamorous biting her nails*", were used to explain the task. This was followed by a practice session in which the participants were asked to provide grammaticality judgments to 6 practice sentences and feedback to their responses were provide. Participants were instructed to give grammaticality judgment as soon as they saw a test prompt on the computer screen. Before each sentence started, a fixation cross was presented for 150 msec in the center of the screen. A 50 msec silence was inserted between words. This allowed for relatively natural-sounding sentence material and no coarticulation across word boundaries. After the final word of the sentence was presented, a 1000 msec pause occurred followed by a test prompt asking the participants to give a button response regarding the sentence's grammaticality. After a response was provided, a black screen appeared and participants pressed the space bar to listen to the

next sentence. Once the participants finished the practice sentences, they proceeded to the first test block.

3.2.4 Participants

Participants who mastered the AGL task by achieving grammatical judgment scores of 90% or greater took part in this task.

3.2.5 Stimulus Materials

A total of 1,102 sentences were used in this task. This included the original 480 experimental sentences in Osterhout et al. study, 480 corresponding filler sentences, additional 71 experimental sentences for the three new verbs based on the results of the corpus analyses, and 71 corresponding filler sentences. These sentences were randomly assigned to each of the 12 blocks, 92 sentences for each of the first 11 blocks and 90 for the last block.

3.2.6 ERP Data Acquisition

ERP recording and averaging were performed in the same manner as in the AGL task (see 3.1.6). In the NLSP task, ERPs were time-locked to the onset of the verb in the sentential complement clauses (e.g., "*was*" in "*The captain believed the crew was unhappy*"). To provide a good baseline for the ERP analyses, the four verbs repetitively used in the sentential complete clauses ("*was*", "*would*", "*had*", "*were*") that were time-locked to have the same duration of 566.6 msec.

3.3 Overall Procedure

Figure B6 summarized the overall procedure of the study. In day 1, participants took part in the pre-training test of the AGL task and were asked to provide grammaticality judgments by guessing. In addition to grammaticality judgments, ERPs were measured while they were performing the task. The participants came back in day 2 and received the first training session of the artificial grammar. At the end of the training

session, a short test of 24 questions was given. The performance was then evaluated based on a criterion of 90% accuracy to determine if one required more training sessions. Each participant was given up to 4 training sessions (on separate days). Once the criterion was reached, the participants came back within 2 days for the post-training test, during which both grammaticality judgments and ERPs were measured. Participation discontinued if one failed to reach the criterion within 4 training sessions. The NLSP task was conducted after the post-training test of the AGL task. Therefore, grammaticality judgments and ERP data of the NLSP task were only obtained from those who reached 90% accuracy in learning the artificial grammar.

3.4 Predictions

Empirical studies have provided ample evidence suggesting that language acquisition, at various levels, can be achieved by statistical learning (Altmann, 2002 b; Chambers, Onishi, & Fisher, 2003; Christiansen, & Chater, 2004; Gerken, Wilson, & Lewis, 2005; Gómez, 2002; Gómez & Gerken, 1999; Gómez & Lakusta, 2004; Onnis, Gómez & Gerken, 1999; Saffran et al., 1996; Saffran, 2001a, 2001b, 2002; Yu & Smith, 2007). However, whether or not statistical learning is at work in language acquisition is still open to debate as grammatical acquisition is considered by some researchers as algebraic learning of symbolic rules as statistical learning is incapable of learning abstract linguistic structures (Marcus et al., 1999; Marcus, 2001; 2003; Peña et al., 2002). This approach is compatible with the conventional generative/minimalist approach to grammatical acquisition.

Figure B7 summarized the predictions of the two contrastive approaches in terms of the mean amplitude of the P600 component. If grammatical learning involves solely learning symbolic rules while disregarding the statistical structure of the grammar, then we would expect to see categorical P600 responses: conditions containing grammatically correct sentences, including Excellent Fit, Good Fit, and Low Fit conditions, would show

no P600 effect, and the condition containing ungrammatical sentences, the Bad Fit condition, would show P600 effects. Contrastively, if the learning system does make use of statistical information in learning the grammar, we would expect to see statistically-based graded P600 effect across condition, with the largest P600 appearing in the Bad Fit condition, which is followed by the low fit condition, which is in turn followed by the Good Fit condition and then the Excellent fit condition.

The same patterns of predictions also held true for the NLSP task. It is noted that in Osterhout et al. study, the P600 magnitude differences between the Excellent Fit (intransitive verbs) and Good Fit (intransitively-biased verbs) did not reach significant. It is possible that the effect was diminished with the original subcategorizations based on results from a sentence completion task (Connine et al., 1984). The new grouping based on results from the corpus analyses therefore provided new categories of verbs with more homogeneous DO/SC distributions.

Finally, it is predicted that the same neural mechanisms were recruited for both processing natural language sentence stimuli and sentences generated by the artificial grammar. This could be indicated by no topographic differences of the P600 effects would be found between the AGL task and the NLSP processing, and this “no effect” is not due to limited power in statistical analyses.

CHAPTER IV

BEHAVIORAL AND ERP DATA

In this chapter we will report results of grammaticality judgments and ERP responses directly addressed the research questions we asked in this thesis. In 4.1, we will focus on the grammaticality judgments in the AGL and the NLSP task. Three comparisons were carried out based on the results of this behavioral measure. First of all, performance in the last training session and the post-training session (1-2 days after the last training session was compared) was compared to examine if our participants showed good retention of the knowledge. In addition, we examined the accuracy of grammaticality judgments in the pre-training test. Performance around a chance level would indicate that the materials used in the AGL task did not contain properties that would bias the performance. Finally, we focused on the behavioral results in the NLSP task. High accuracy in grammaticality judgments would confirm that the participants understood the task and were able to provide adequate grammaticality judgments. The results from grammaticality judgments were then used to filer ERP trails such that only trails with a correct grammaticality judgment were included in subsequent ERP data analyses. The purpose of doing so was to code with potential mixture responses discussed in 3.2.1. In 4.2, we examined ERP responses of the AGL task. Specifically, we compared results from the pre- and post-training session to examine whether P600 effects associate with the experience or the state of knowledge of the learners. In addition, we also examined whether the magnitude of the P600 varies systematically with the conditional probability of the input after learning. This was done by comparing the magnitude of

P600 effects across conditions in the post-training session of the AGL Task. The same comparisons were conducted to examine magnitude differences among conditions for the NLSP task. The results were reported in section 4.3. In section 4.4, we further investigated whether P600 effects elicited by syntactic processing of natural language sentences conform to that elicited by processing sequential patterns learned in a statistical learning task. To this end, the difference waves were calculated by subtracting the mean voltage of the Excellent Fit condition (baseline) from each of the rest three conditions for both tasks. In 4.4.1, we reported results from comparisons of the ERPs responses for the AGL and the NLSP tasks using difference waves. In 4.4.2, the same comparisons were carried out by including additional 8 nearby electrode sites and using smaller latency windows. The purpose of using extended locations and smaller latency windows was to inspect potential differences between the two tasks.

4.1 Grammaticality Judgments

14 participants out of 23 reached 90% accuracy within 4 or less training sessions. Among the 14 participants, data from two of them were later excluded from data analyses because more than 30% of the experimental (ERP) trials were contaminated due to an excessive number of eye blinks/movements or poor data quality. The results reported in this chapter were based on data from the 12 participants (age range = 20-41 years).

Table A5 shows the means and standard deviations of the grammaticality judgments in the NLSP task and the pre- and post-training tests of the AGL task. For the post-training test in the AGL task, all 4 conditions showed above 90% accuracy, indicating good retention of knowledge from the last training session (mean accuracy = 96.33, *SD* =2.97). Mean accuracy differences among the 4 conditions in post-training test were not significant, F (3, 33) = 1.42, p = .254.

To assure that the expected ERP differences were actually caused by differences in syntactic processing, before our participants received any training they gave

grammaticality judgments to the same set of test sentences used in the post-training test. Overall, the mean accuracy for each condition in the pre-training test was close to a chance level of 50% (see Table A5). This is statistically confirmed with a non-significant difference between the overall accuracy (pooling across the 4 conditions) of the pre-training test and the chance level, t (1, 11) = -.44, p = .67. Improvement in performance was indicated by significantly higher mean accuracy in the post-training test than the pre-training test for all four conditions (t_{11} = 21.94, p < .0001 for the Excellent Fit condition; t_{11} = 14.40, p < .0001 for the Good Fit condition, t_{11} = 18.34, p <.0001 for the Low Fit condition, t_{11} = 24.08, p < .0001 for the Bad Fit condition).

For the NLSP task (see also Table A5), high levels of grammaticality judgments were also observed in both the original grouping and the new grouping of the verbs, indicating that the participants were able to provide adequate grammaticality judgments for the sentences. Mean accuracy differences among the 4 conditions were not significant in terms of either the original grouping, F (3, 33) = .85, p = .48, or the new grouping, F (3, 33) = .74, p = .54.

4.2 ERPs: AGL Pre- and Post-training Tests

Given that the significant improvement of performance in grammaticality judgments from pre- to post- training tests is not likely due to some specific properties of the words but is reflective of learning, we can then turn to examine whether the P600 effect associates with the experience or the state of knowledge of the learners. To this end, ERPs at Pz and within 500-800 window, the typical electrode site and latency window for the P600 component, in the pre- and post-training test of the AGL task were compared. Figure B8 displays the ERP responses for each condition at Pz within 500-800 msec window in the pre-training test. The mean voltages and standard deviations for each condition in the pre- and post-training tests are shown in Table A6. ERPs at the same electrode site and within the same latency window in the post-training test are shown in

Figure B9. Comparisons of the two figures indicate a clear distinction of the ERPs obtained in the pre- and post-training. In the pre-training test, the mean voltage within the 500-800 mesc window was close to zero for all 4 conditions. However, in the post-training test positive-going waveforms were observed in the same window. Critically, the mean voltages of the 4 conditions diverged, revealing graded P600 effects.

A Test Phase (Pre- and Post-Training Test) x Conditions (Excellent, Good, Low, and Bad fit) repeated measure analysis of variance (ANOVA) was carried out to examine the observed results. A significant main effect of test phase was found, F (1, 11) = 11.03, p =.007, η^2 = .50, revealing an overall larger mean voltage of the P600 in the post-training test than the pre-training test. The main effect of conditions also reached significance, F (1, 11) = 3.97, p = .016, η^2 = .27, which may be attributable to the interaction between test phase and condition, F (3, 33) = 9.54, p <.0001, η^2 = .47. Step-down repeated measure ANOVAs indicates a significant effect of condition for the post-training test, F (3, 33) = 27.31, p <.0001, η^2 = .71. Post-hoc pair-wise comparisons were then carried out to examine whether the magnitude of the P600 varies systematically with the conditional probability of the input after learning. The results showed graded mean amplitude of the P600 across conditions: the Bad Fit condition elicited the most positive P600 effects among all 4 conditions (Bad Fit > Low Fit: t_{11} = 3.27, p = .007 ; Bad Fit > Good Fit: t_{11} = 4.94, p < .0001; Bad Fit > Excellent Fit: t_{11} = 7.16, p < .0001), which is followed by the Low Fit condition (Low Fit > Good Fit: t_{11} = 3.40, p = .006; Low Fit > Excellent Fit: t_{11} = 5.75, p < .0001), which is in turn followed by the Good Fit condition (Good Fit > Excellent Fit: t_{11} = 3.33, p = .007) and finally the Excellent Fit condition. The results reveal a systematic variation of the magnitude of the P600 and the goodness of fit of materials determined by input statistics acquired during training. For the pre-training test, differences among conditions were not significant, F (3, 33) = .78, p = .52, η^2 = .07. The lack of P600 effects before training together with the emergence of graded P600

effects once the learners possessed a high level of knowledge about the materials suggest that statistical learning is at work in learning the artificial grammar.

4.3 ERPs: NLSP Original Grouping and New Grouping

Given that the magnitude of the P600 varied systematically with the conditional probability of the input after learning in the AGL task, we further examined if the same patterns exists in processing natural language**.** We first focused on the original grouping of the verbs. Figure B10 displays the grand average ERPs at Pz within 500-800 msec window for each condition. The mean voltages and standard deviations are shown in Table A7.

An inspection of Figure10 reveals a similar pattern of P600 effects across conditions seen in the post-training test of the AGL task (see Figure B9). A one-way repeated measure ANOVA revealed a significant effect of condition in the NLSP task, $F(3, 33) = 29.75$, $p < .0001$, $\eta^2 = .73$. Post-hoc tests confirmed that the magnitude of the P600 varies systematically with the conditional probabilities of the input in processing natural language. The sentences in the Bad Fit (transitive verbs) condition elicited the largest P600 effects among the 4 conditions, which is followed by the Low Fit (transitively-biased verbs) condition, which is in turn followed by the Good Fit (intransitively-biased verbs) condition. Different from the Osterhout et al. study, we found a significant difference between the Good Fit condition and the Excellent Fit condition. This difference might be due to the use of a balanced amount of target and filler sentences and also the use of behavioral measure to filter out EEG trails with an incorrect response, two major modifications we made to form the current version of the task used in this thesis (see 3.2.1).

Parallel analysis was conducted to examine mean amplitude differences among conditions using the new grouping. Figure B11 shows the grand average ERPs at Pz within 500-800 msec latency window for each condition. The corresponding mean

voltages and standard deviations are provided in Table A7. A statistical analysis indicates a significant effect of condition, F (3, 33) = 25.67, p <.0001, η^2 = .70. Post-hoc comparisons revealed graded P600 effects across conditions: Bad Fit > Low Fit > Good Fit (Bad Fit > Low Fit: t_{11} = 3.83, p = .003 ; Bad Fit > Good Fit: t_{11} = 4.82, p = .001; Bad Fit > Excellent Fit: t_{11} = 7.93, p < .0001; Low Fit > Good Fit: t_{11} = 3.37, p = .006; Low Fit > Excellent Fit: t_{11} = 5.64, p < .0001; Good Fit > Excellent Fit: t_{11} = 3.52, p = .005). However, with the new grouping of verbs, the difference between the Good Fit and Excellent Fit conditions became non-significant (t_{11} = -1.08, p =.306), although differences between the Bad Fit and the Excellent Fit conditions and between the Low Fit and the Excellent Fit conditions were still significantly (p <.0001 for both comparisons). Given the results, subsequent analyses of the NLSP task focused only on data with the original grouping. More discussion on the results with the new grouping was provided in Chapter 5.

4.4 Comparisons of the AGL and NLSP Tasks

In section 4.2 and 4.3, we found the same patterns of graded P600 effects across conditions that varied systematically with the statistical patterns of the input in both tasks. To further examine whether the P600 obtained in the AGL task conform to the P600 obtained in the natural language task, we compared the P600 effects in the two tasks directly. The results would provide information to whether the same neural mechanisms were recruited for both syntactic processing of natural language and statistical learning of sequential patterns in an artificial grammar (Christiansen, Conway Onnis, 2007; Patel et al., 1998). Difference waves were calculated for Conditions Good-Excellent, Low-Excellent, and Bad-Excellent for both tasks. The results are summarized in Table A8. By using difference waves, task-specific effects such as different vocabularies are removed, leaving only those effects due to differences among conditions.

Two sets of analyses were conducted based on difference waves. In 4.4.1, a Task (AGL, NLSP) x Condition (Good-Excellent, Low-Excellent, Bad-Excellent) repeated-measure ANOVA was conducted to examine potential between task differences of the P600 effects. Critical to the question of interest is whether the task by condition interaction is significant. A significant interaction between task and condition would reject the hypothesis that processing sentences or strings in the two tasks was based on the same neural mechanisms.

In 4.4.2, we further examined the same question by using smaller latency windows and including more electrodes in addition to the typical electrode site (Pz) of the P600. First of all, we sliced the 500-800 msec latency window into 3 smaller windows: 500-600 msec, 600-700 msec, and 700-800 msec. The use of smaller windows allows us to detect fine differences between the two tasks. Furthermore, we included additional 8 surrounding electrodes, including C3, CZ, C4, P3, PZ, P4, O1, OZ, O2, of the Pz in the analysis. The rationale behind using more electrodes in our analysis is to take advantage of spreading voltage distributions at the scalp, which is in fact often considered a disadvantage of the ERP technique. Theoretically, it is not impossible for two distinct neural generators to give rise to the same ERP responses measured at a given electrode site at the scalp. This has to do with the fact that when a dipole is present in a conductive medium such as the brain, electricity does not just run directly between the two poles of a dipole, but instead spreads out through the conductor (Luck, 2005). ERPs spread out as they travel through the brain following the path of least resistance. Consequently, an ERP generated in one part of the brain can lead to substantial voltages at rather distant parts of the scalp (Luck, 2005), which could give rise to similar ERP responses measured at some locations even if they come from different sources. This is relevant to the so-called inverse problem for even information about voltage distribution at the scalp is available; it is unlikely to identify the locations and orientations of the generator dipoles (the *inverse problem*). Instead of trying to deal with this problem, we took advantage of the

spreading voltage at the scalp by including more electrodes in our analysis. The idea behind this was that if the same neural mechanisms were recruited in syntactic processing of the AGL and NLSP tasks, the voltage distribution at the scalp elicited by the stimuli used in the two tasks should be the same with the latency window for the P600 component.

4.4.1 P600 at Pz within 500-800 msec Latency Window

Figure B12 and B13 show the difference waveforms for the Good Fit, Low Fit, and Bad Fit conditions in the AGL and the NLSP task, respectively. A Task (AGL, NLSP) x Condition (Good Fit, Low Fit, Bad Fit) repeated-measure ANOVA was conducted to examine potential differences between the two tasks. Most relevant to the question of interest is the interaction between task and condition. A significant interaction would reject the hypothesis that the same neural mechanisms were recruited for syntactic processing of the two tasks. Results from the statistical analyses indicated a non-significant task by condition interaction, F (2, 22) = .72, p = .49, η^2 = .06, suggesting no evident differences among the two tasks. The main effect of task was not significant, F (1. 11) = 2.50, p = .14, η^2 = .19. A significant main effect of condition was found, F (2, 22) = 44.25, p < .0001, η^2 = .79. Follow-up comparisons by pooling across the two tasks revealed the same pattern of magnitude differences of the P600 across conditions observed in previous analyses: Bad Fit > Good Fit > Low Fit> Excellent Fit (p <.0001 for all pair-wise comparisons).

4.4.2 P600: Extended Electrode Sites

To more closely compare the P600 effects in the AGL and the NLSP task, the 500-800 msec latency window was sliced into 3 smaller windows with an equal duration of 100 msec: 500-600 msec, 600-700 msec, and 700-800 msec. Nine electrodes, C3, CZ, C4, P3, PZ, P4, O1, OZ, O2, in the central-parietal region were included in this analysis. The ERP responses in the Good Fit-Excellent Fit, Low Fit-Excellent Fit, and Bad Fit-

Excellent Fit conditions for each of the 9 electrode sites within each of the 3 latency windows are shown in Figure B14 to Figure B16 for the AGL task and Figure B17 to Figure B19 for the NLSP task. Corresponding mean amplitudes and standard deviations are summarized in Table A9 for the AGL task and Table A10 for the NLSP task.

Repeated measure ANOVAs by including task (AGL, NLSP), conditions (Good-Excellent, Low-Excellent, and Bad-Excellent), and the 9 electrode sites as 3 within-subject factors was conducted once for each of the three windows. Since the purpose of this set of analyses was to further inspect potential differences of the P600 effects between the two tasks, the following report of the statistical results was organized in terms of (1) *Were there between task differences within each of the 3 windows*? This question was examined by focusing on the main effect of task and the effect of interaction between tasks and conditions for each of the three latency windows; (2) *Were there scalp distribution differences between the two tasks within each of the 3 windows*?

First of all, we examined the interaction between task and condition for each of the three windows. The main effect of tasks was not significant for any of the 3 latency windows (500-600 msec: F (1, 11) = 1.13, p =.31, η^2=.93; 600-700 msec: F (1, 11) <.0001, p = .91, η^2 <.001; 700-800 msec: F (1, 11) =.029, p =.87, η^2 =.003). For the Task x Condition interaction, again the effects were not significant for any of the three latency windows (500-600 msec: F (2, 22) = .19, p =.83, η^2=.17; 600-700 msec: F (1, 11) =.03, p = .97, η^2 =.002; 700-800 msec: F (2, 22) =.81, p =.23, η^2 =.13). These results are consistent with the finding of non-significant task differences from the previous analysis using the whole 500-800 latency window.

We also examined potential differences on scalp distribution for the AGL and the NLSP task by focusing on interaction effects between task and electrode sites and/or conditions. First of all, we focused on the two-way interaction between task and electrode sites. The Task x Electrode interaction was not significant for any one of the three latency windows (500-600 msec: F (8, 88) = 1.41, p =.21, η2=.11; 600-700 msec: F (8, 88) = .62,

1.13, p = .76, $\eta 2$ =.05; 700-800 msec: F (8, 88) =.81, p =.60, $\eta 2$ =.07). Similarly, none of the three-way interaction among task, electrode sites, and conditions reached significance (500-600 msec: F (16, 176) = .71, p =.78, $\eta 2$=.06; 600-700 msec: F (8, 88) = .71, p = .78, $\eta 2$ =.06; 700-800 msec: F (8, 88) =.74, p =.75, $\eta 2$ =.06).

In sum, the results from this set of analyses indicated no evidence of differences between the two tasks on scalp distributions. The statistically undistinguishable scalp distributions provides one piece of evidence suggesting that the same neural mechanism might have been recruited for both processing the natural language stimuli and the sentence strings in the AGL task.

CHAPTER V

LOOKING BEYOND P600

In chapter 4, we will report results directly addressed the research questions asked in this thesis. In the AGL task, more than one third of the total amount of the participants recruited failed to reach the criterion in learning the artificial language. This rather high attrition rate raised a question about the extent to which our results are compatible to previous studies on AGL. In section 5.1, we will further examine our participants' learning performance across training sessions and compare our results with previous studies.

Before the NLSP task was conducted, we performed a corpus analysis to confirm the verb categorizations in the study by Osterhout and colleagues. However, the new grouping of verbs based on the results of DO/SC probability in the corpus analysis did not give rise to expected graded P600 responses. In section 5.2, we will further explore this issue and possible interpretations will be provided. In section 5.3, we will further compared ERPs elicited in the AGL and the NLSP tasks, by including other ERP time windows. It is noted that using different time frames in ERP measures in principle mean looking at different ERP components. This is different from the analyses we reported in 4.4.2 in which the overall time frame still spanned over the typical 500-800 msec time frame of the P600 but was sliced into 3 smaller windows. Therefore, the analyses we performed in 5.5 were to further examine *overall* between task differences, rather than just focusing on the P600 component. Although the focus of this thesis was on the grammatical learning/processing and therefore the major ERP analyses were conducted on the P600 component, the data from other ERP time windows will be examined to

determine whether the effects of probabilistic variation in the naturalistic or AGL stimuli result in different ERP responses.

In chapter 4, we compared the P600 effects obtained in the AGL task conform to the P600 obtained in the natural language task and did not find differences between the two tasks, suggesting similar processes involved in syntactic processing of sentences in the two tasks. It is also of interest to compare other aspects of the two tasks given there are apparent between task differences such as a lack of semantics in the AGL task. In section 5.3, ERPs in the two tasks were compared using latency windows not only specific for syntactic processing (500-800 msec) but also latency windows in which other language related ERP components, such as N400 and ELAN/LAN, have been reported to occur.

5.1 Individual Difference in Learning the Artificial Grammar

In the AGL task, 9 participants failed to reach 90% of accuracy in learning the artificial grammar within 4 training sessions. The constituted about 39% of the total amount of participants recruited (n = 23). The rather high failure rate raised a concern about the specific artificial grammar used in this thesis and a question about individual differences in such learning. In this section, we focused our discussion on these two aspects and further compared our participants' performance with previous studies on AGL.

Figure B20 shows the mean percentage correct in the test given at the end of each training session for all 23 participants. In Figure B20, each line corresponds to each participant. Note that the number of scores each participant had depends on how the participant progresses through the training. For instance, one of the participants only has a single data point because he scored 97.5% in the first learning session.

To compare the current results with previous work on AGL, Table A11 summarized some previous studies on AGL in adult learners. All of these studies involved a finite-state grammar, despite the types of stimuli (novel words or letter strings), modalities of stimuli presentation, duration of training, and test paradigms in these studies differ. Most of these studies involved 1-2 short training sessions, although specific training duration was not always available. Therefore, it is sensible to compare our participants' performance at the end of the first training session with the results in these studies.

An inspection of Figure B20 indicates that all except for one of our participants showed above 50% accuracy at the end of the first training session. A further analysis indicated significantly above chance level performance at the end of the first training session, $t_{22} = 8.43$, $p < .0001$. The mean accuracy across all participants (n = 23) at the end of the first training session was 75.43 ($SD = 14.46$), which is in fact higher than that in most of the studies summarized in Table A11, except for the study by Friederici, Steinhauer, and Pfeifer (2002) and the one by Christiansen, Conway, and Onnis (2007). Different from the rest of the studies summarized in Table A11, Friederici and colleagues used a pre-set criterion to determine the amount of training for individual participants. This is similar to the use of a 90% accuracy criterion in the current study. Their participants were trained until a 95% accuracy criterion was reached. Therefore, the high level of performance reported by Friederici and colleagues was a result of more training received. Therefore, a compatible comparison with their study would therefore be the post-training test rather than the first training session in our task. For the study by Christiansen, although the training time was much shorter, the data reported in their study were based on those who reached a high level of performance in learning the artificial language (Christiansen, person conversation). In general, the learning performance in the current study performance is compatible with the results reported in previous studies

using a finite state grammar. The seemingly high failure rate in the current study has to do with the relatively higher criterion (90%) used to determine performance.

For participants who did not reach the 90% accuracy criterion, there is no evidence that they did not learn the materials as they showed increased accuracy across training session. The slightly better overall performance in the current study might be contributable to prior exposure of the stimuli in the pre-training test before the first training session started. In the pre-training test session our participants were informed that half of the test trails were ungrammatical and were asked to give grammaticality judgments by just guessing. Although the participants were naive to the regularities of the grammar as their performance was around the chance level in the pre-training test, the participants were left with a great amount of experience with words from the artificial grammar, which could result in faster learning of the materials once the training started.

5.2 The New Verb Grouping in the NLSP Task

Two different verb groupings, one followed the original grouping in the study by Osterhout and colleagues and the other followed the results in the current corpus analyses, were used in the NLSP task. When the original grouping was used, we obtained similar graded P600 responses across conditions. In addition, the difference between the Excellent Fit (intransitive) and Good Fit (intransitively-biased) conditions, which was not significant in the study by Osterhout and colleagues, also reached significance in the current study. This additional significant effect found in this these might be attributable to two of the modifications we made to their task: 1) having an equal amount of filler sentences as the target sentences such that the target sentences were examined under a more neutral context; 2) using behavioral results to filter out ERP trails included in data analyses.

The new verb grouping was based on results from the corpus analyses. In the new grouping, one major change from the original grouping was that the mean SC probability

difference between the Excellent Fit and the Good Fit conditions became smaller (94.46-57.65%=36.81 in the original grouping; 96.19%-70.03%= 26.16% in the new grouping). Therefore, if the mean amplitude of the P600 is sensitive to the goodness of fit between already heard materials and subsequent elements, we would see a corresponding smaller difference between the two conditions in terms of the magnitude of the P600 in the new grouping. This is confirmed by comparing the difference between the Excellent Fit (blue line) and the Good Fit (green line) conditions in Figure B10 (the original grouping) and B11 (the new grouping). However, when the new grouping was applied, the magnitude differences of P600 between the Excellent Fit and Good Fit conditions became insignificant.

There are two possible interpretations to this effect. First of all, the nonsignificant effect suggests that "verb transitivity" might better capture the state of the knowledge better than DO/SC categories. However, since the 3 corpuses used in the current study together were composed of only 3 million words, whether verb transitivity or SC/DO probabilities is more representative of the state of the knowledge should be further evaluated with corpuses of a larger sample size.

Perhaps a more sensible interpretation is that the relationship between input statistics and neurophysiological responses is not linear. This is often seen in work on computational modeling of neural networks in which communication between individual processing elements (neurons) are based on certain transfer function, typically non-linear, that generates a single output value from all of the input values that are applied to the neuron. When the DO/SC probability difference between the Excellent Fit and the Good Fit conditions was adjusted in the new grouping, differences in neurophysiological responses became too subtle to show in ERP measures. After all, the P600 effect of the Good Fit condition in the new grouping did moved following a predicted direction (toward the Excellent Fit condition), suggesting that the magnitude of P600 effects is sensitive to this change. The use of the new grouping in fact increased the mean SC

probability difference between the Good Fit and the Low Fit conditions (from 41.53% in the original grouping to 44.66% in the new grouping) and between the Low Fit and Bad Fit conditions (from 15.94% in the original grouping to 25.19% in the new grouping), the remaining significant differences between conditions in terms of the P600 was expected.

5.3 Overall Comparisons of ERPs in the AGL and NLSP Tasks

Following the previous set of analyses, in this section we further compared the two tasks by expanding the time and location parameters from the previous set of analyses. Four windows with a latency of 300 msec used in this set of analyses were: 100-400 msec, 300-600 msec, 500-800 msec, and 700-1000 msec. Different from Analysis II using electrode sites surrounding Pz, the electrode sites included in this analyses constituted a more spread-out scalp region. These included: left anterior (F3/FC3), frontal central (FZ/FCZ), right anterior (F4/FC4), left central (C3/CP3), central (CZ/CPZ), right central (C4/CP4), left posterior (P3/02), central posterior (PZ/OZ), and right posterior (P1/02). With the 4 latency windows and expanded scalp locations, it is possible to capture not only potential differences of the P600 component for the AGL and the NLSP task, but also other components associated with task-specific characteristics. Repeated measure ANOVAs with task (AGL, NLSP), condition (Good-Excellent, Low-Excellent, and Bad-Excellent), and electrode sites as 3 within-subject factors was conducted once for each of the 4 windows. Overall, the results for the 300-600 msec, 500-800 msec, and 700-1000 latency windows are similar to that in Analysis III. Neither the main effect of task nor the interaction between task and condition was significant for the three windows. However, there was a significant interaction between task and electrode sites within the 100-400 msec latency window, F (8, 88) = 2.03, p = .05, η^2= .16. This was further confirmed with a significant main effect of task, F (1, 11) = 9.23, p = .01, η^2= .46. The Task x Condition x Electrode site interaction was not

significant. Although the significant task by electrode site interaction in this latency window do not reflect the principal ERP component of the current study (i.e., P600), it is also of interest to locate potential differences between the two tasks.

To obtain further information regarding where the interaction effects came from, topographic maps based on difference wave were plotted for the AGL task (Figure B21) and the NLSP task (Figure B22). Visual inspection of the figures between the two tasks revealed ERP negativity distributed in the frontal scalp area only for the NLSP task, but not the AGL task. To confirm this observation, statistical comparisons between the two tasks and the 3 difference wave conditions were conducted by collapsing across the three frontal sites. A significant main effect of task was found, F (1, 11) = 8.13, p = 0.16, η^2= .46, with the NLSP task showing a more negative mean voltage than the AGL task within the 100-400 msec latency window. The main effect of condition and the interaction between condition and task did not reach significance.

The observed negativities within the 100-400 msec latency window is reminiscent of N400 and ELAN/LAN (early left anterior negativity/left anterior negativity). These are language-associated ERP components associated with different aspects of linguistic processing (Hahne & Friederici, 1999; Hagoort, Brown, & Osterhout, 2000). N400 is usually largest over centro-parietal scalp sites and is associated with semantic fit of a word in a given sentence context (Kutas & Hilyard, 1980; Kutas & Van Petten, 1994). Different from N400, ELAN/LAN effects are often observed over left anterior scalp sites (Hahne & Friederic, 1999). The left anterior negativities have been observed in association with phrase structure violation (Friederici et al., 1993, 1996; Münte, Heinze, & Mangun, 1993; Neville et al., 1991; Osterhout & Holcomb, 1992, 1993), with the processing of subcategorization information (Osterhout & Holcomb, 1993; Rösler, Friederici, Pütz, & Hahne, 1993), and with agreement violations (Coulson et al., 1998; Friederici et al., 1993;Gunter et al., 1997; Osterhout & Mobley, 1995). Moreover, a left anterior negativity was observed for the processing of closed-class word (Neville, Mills,

& Lawson, 1992; Nobre & McCarthy,1994). Given that the negativities in the NLSP task have a frontal scalp distribution and that it started relative early (see Figure B22), it is likely that the observed negativities are reflective of ELAN/LAN effects. Friederici and colleagues (Hahne & Friederici, 1999; Friederici, 2002) suggested that ELAN and LAN are two distinguishable components; the former has a latency of about 100 to 300 msec and the latter with latency of about 300-500 msec. ELAN has been observed for processing of phrase structure violations and closed-class elements, whereas the LAN is often elicited by violations of subcategorization and inflectional morphology.

As for the early negativities found in the NLSP task, it is likely that the effects are reflective of processing close-class elements in that the ERPs in the NLSP task were time-locked to close-class words, such as “was”, “would”, and “had”. Given that the AGL task is a “miniature” language that contains information no more than sequential patterns of nonsense words, it is certainly not surprising that the ERPs elicited by processing natural language would show effects absent in the EPRs elicited y processing sentence strings generated by the AGL.

CHAPTER VI

GENERAL DISCUSSION

In this study we attempted to extend the empirical evidence that provide the basis for probabilistic accounts of statistical language learning and processing. Traditionally, statistical learning in the classical associative sense has been intimately associated with behaviorism in which learning language is considered no different from learning other behaviors by establishing stimulus-response association (Palmer, 1981). This simple view of statistical learning has been downplayed as it has limited power for learning materials with complex structures. In language in particular, the domination of the rationalist position is a clear indication of resistance to consider statistical learning as a viable approach to language acquisition. Two major issues led to a paradigm shift in language research during the 1960s. These issues centered around whether there is a lack of sufficient amount of information necessary for statistical learning of language (the POS argument), and whether or not children are equipped with strong enough (statistical) learning capacities to achieve language acquisition. These two issues have been addressed with accumulating evidence showing that the language input children hear is abundant in statistical regularities (e.g., Kelly 1992; Kelly & Martin, 1994; Billamn & Knutsen, 1996), and that children are equipped with strong data mining abilities to detect distributional regularities in the input and learn linguistic categories and structures at various levels (e.g., Maye, Werker, & Gerken, 2002; Saffran, Newport, & Aslin, 1996a; 1996b; ; Saffran, Aslin, Johnson, & Newport, 1999; Gómez & Gerken, 1999; Altmann, 2002; Gerken, Wilson, Lewis, 2005).

Although this research has been highly valuable, whether or not statistical learning is at work in learning *grammar* remains a subject of debate in the field of grammatical acquisition Using the AGL paradigm, researchers have reached different conclusions as to mechanisms involved in such learning and the essence of the knowledge learned in AGL tasks. For instance, Marcus and colleagues (1998a, 1998b, 1999, 2001, 2003) argued that infants acquire algebra-like rules that operate on abstract unitary symbols and that systems sensitive only to statistical regularities are in principle incapable of such abstraction. Using similar tasks, others have argued for a robust role of statistical learning. This reflects a modern variant of the long standing debate between nativists and empiricists/interactionists account where learning plays a central role but does so within biological constraints (Gómez &Gerken, 1999; Onnis, Christiansen, & Chater, 2004). Some researchers have argued for a middle ground (Bonatti et al., 2005; Peña et al., 2002), but such a compromise has not really resolved the debate in grammatical acquisition as statistical learning is still assumed to be at work *before* grammatical acquisition starts (Keidel et al., 2007; Onnis et al., 2005; Seidenberg et al., 2002). A further question following this debate has been the extent to which results obtained in AGL task can be applied to natural language acquisition.

In this thesis we focused on the area of grammar and examined whether or not statistical learning plays a role in syntactic learning and processing. We contrasted the statistical learning account of grammar with grammatical acquisition theories taking the most extreme version of the nativist approach to language acquisition (Chomsky, 1965; Wexler & Culicover, 1980). Many accounts of grammar assume that at least by adulthood, the grammatical representations are fully abstract such that one should not expect to see grammatical representations that are probabilistic and thus graded. Adults or learners with mature language knowledge are assumed to process language based on abstract symbolic categories such as *subject*, *object*, or *verb* as proposed in the nativist account or symbolic categories such as X_ _X_ X as claimed by Marcus and colleagues

if an artificial language is learned. If evidence of statistical basis for grammatical learning is found in adult participants, this evidence should carry considerable weight toward a view of grammar throughout development as representations comprised of probabilistic statistics. For this reason, we focused on adults' grammatical learning and processing using ERP measures.

6.1 Empirical Tests of the Statistical Learning Account of Grammatical Learning

The goal of this thesis was to examine the statistical learning account of grammatical acquisition. We also sought to validate the use of AGL to study natural language grammatical acquisition. These goals were accomplished by collecting complementary data from a natural language sentence processing (NLSP) task and artificial grammar learning (AGL) task.

6.1.1. Statistical Learning of Artificial Grammar

The goal for collecting data using the AGL task was threefold. First, we included a pre-training test phase in which our participants gave grammaticality judgments to the same set of test questions used in the post-training test. Electroencephalography (*EEG*) was also recorded during the post-training phase in addition to the behavioral measure. Previous studies in the literature have raised a concern that prior language experience might influence performance on learning an artificial grammar (Onnis, Monaghan, & Chater, 2005). The inclusion of a pre-training test phase was to assure that the expected ERP differences were actually caused by learning. Second, we compared ERP responses between the pre- and post-training sessions. Specifically, we examined whether the P600 effect associates with the experience or the state of knowledge of learners. We predicted that P600 effects would be seen once the grammar is learned to a high level (post-training phase), which contrasts with no P600 effect before training. In addition, we predicted that the magnitude of the P600 would vary systematically with the conditional probabilities of

the input or “goodness of fit” of a word with already unfolded materials after learning, which such systematic variations would not be seen before learning.

Results from the grammaticality judgments and ERP responses support our predictions. First of all, the learners’ accuracy in grammaticality judgments was around the chance level in the pre-training phase. This is coupled by above 90% accuracy in grammaticality judgments in the post-training test. Since the post-training session was at least one day apart from the last training session, the high level of grammaticality judgments indicates that our adult learners were excellent at retaining the grammatical knowledge. Comparisons of ERP responses in the pre- and post-training phases showed emergence of P600 effects as a function of learning, supporting the prediction that the P600 effect associates with the experience or the state of knowledge of learners. Critically, the magnitude of the P600 in the post-training phase varied systematically with the “goodness of fit” defined by conditional probability of a word given the preceding sentence material in the input. The possibility that the graded P600 effects across condition was due to mixture of responses was reduced, if not fully eliminated, by including only trials that were correctly judged to be grammatical or ungrammatical into ERP data analyses. Along with the finding of no difference among conditions in the pre-training test phase, the results suggest sensitivity of the learning mechanisms to statistical properties of the materials in learning grammar.

Taken together, the data from the AGL task suggest a critical role of statistics in learning the grammar of an artificial language. The fact that the influence of input statistics was still well-captured in processing sentence strings once the artificial grammar was learned to a high level suggests that input statistics not only play a *bootstrapping* role in grammatical learning, but also have moment by moment influence on language processing (use). It is noted that these findings do not totally eliminate the possibility that abstract categories also develop over time. However, the graded, rather than categorical, P600 effects seen in the post-training phase refutes the argument that

statistical learning is incapable of learning grammar, such as that represented in the AGL task.

6.1.2 Natural Language Processing

Grammaticality learning has traditionally been conceived of as learning and processing of categorical linguistic abstractions. Therefore, it could be argued that the graded, rather than binary, categorical P600 responses we saw in the AGL task (post-training) were due to insufficient training or perhaps that the AGL task was indeed artificial and thus did not invoke the kind of learning mechanisms involved in acquiring a natural language. This concern is especially plausible as we consider the amount of experience our participants have with their native language versus the artificial grammar. The use of the NLSP task provided a chance to further explore this possibility.

Our participants showed high levels of grammaticality judgments in all 4 conditions, indicating that the participants were able to provide adequate grammaticality judgments for the sentences. Critically, we also obtained the same pattern of graded P600 responses, ***Excellent Fit < Good Fit < Low Fit < Bad Fit,*** as was seen in the AGL. Different from the study by Osterhout and colleagues in which the difference between the Excellent Fit (intransitive verb) and Good Fit (intransitively-biased verb) condition was not significant, the difference between the two conditions reached statistical significance. Two critical modifications we made to their task might contribute to the appearance of a significant difference between the 2 conditions. First of all, the NLSP task used in this thesis has a balanced amount of target and filler sentences, while in the original task the number of the target sentences was three times as great as the amount of the filler sentences. In their study, participants heard more sentences with a SC clause. While this particular structure became a predominant one, this stimulus training bias could have neutralized the differences in statistical distributions between the Excellent Fit and the Good Fit conditions. Particularly, the more frequent SC structure could have biased their

participants toward interpreting the post-verbal noun phrase as a *subject noun* for sentences containing a (intransitively-biased) verb in the Good Fit condition. Since post-verbal noun phrases in the Excellent Fit condition always acted as a subject noun of the SC clause, the task-specific structural bias could therefore have diminished ERP differences between the two conditions. In the current version of the task, the number of target and filler sentences was balanced such that 50% of the time our participants heard a sentence in a SC structure and the other 50% of the time in a transitive structure. The balanced amount of target and filler sentences provided a neutral test context such that task-specific factors were controlled, leaving performance variations to effects from prior linguistic experience. A second possible interpretation to the observed difference between their work and the current study has to do with the fact that we filtered out ERP trials with an incorrect grammaticality judgment to cope with potential "*mixture responses*" in ERP data analysis. In their study, acceptability judgments were 91%, 84%, 66%, and 4% for sentences in the Excellent Fit (intransitive verbs), Good Fit (intransitively-biased verbs), Low Fit (transitively-biased verbs), and Bad Fit (transitive verbs) conditions, respectively. Although more sentences in the Excellent Fit condition than the Good Fit condition were judged acceptable, the difference (91% vs. 84%) was not statistically evaluated in their study. By using responses from the grammaticality judgments to filter out sentences that were incorrectly judged, we were able to set comparisons among conditions on a common ground and focused on trials that were responded to correctly. It would be ideal to show that ungrammatical trials with an incorrect grammaticality judgment (i.e., judged as grammatical) did not differ from grammatical trials with a correct grammaticality judgment or grammatical trials with an incorrect grammaticality judgment (i.e., judged as ungrammatical) did not differ from ungrammatical trails with a correct grammaticality judgment. This would provide us some information pertaining to the extent to which mixture responses had influenced the ERP responses in the study by Osterhout and colleagues. However, since the number of such EEG trails was too small to

obtain reliable ERP responses (by averaging), we did not further examine this interpretation in this thesis.

In sum, the results in the NLSP task provided direct evidence of a graded P600 response to grammatical violations using natural language stimuli. These findings suggest that the lack of binary P600 effects seen in the AGL task is not likely due to insufficient learning or the use of an artificial grammar. Even in a task in which the materials have been "*overlearned*" (e.g., the NLSP task) and comprising a natural language, neurophysiological responses vary in a systematic way as input statistical regularities.

6.1.3 Syntactic Processing of Natural Language Sentences and Statistical Learning of Sequential Patterns

Results from the AGL and the NLSP task clearly demonstrated systematic variations between neurophysiological responses and input regularities. Another goal in this thesis was to examine the extent to which we could show that the grammatical learning observed in an AGL task resulted in neural representations that were similar to those formed in natural language grammatical acquisition. To this end, we directly compared ERP responses from the two tasks using difference waves. We found a non-significant interaction between task and condition, suggesting no apparent differences in terms of P600 effects between tasks. To further explore potential differences between the two tasks in terms of P600 effects, we included *the same latency window* of 500-800 msec but sliced it into 3 smaller windows with an equal duration of 100 msec for further analyses. In addition, we expanded the electrode sites to also include an additional 8 nearby electrodes in addition to the typical Pz. Thus, we extended our analysis in space, but within the same time frame with smaller latency windows

This extended analysis provided smaller time windows to examine whether subtle differences in P600 effects could be found between the two tasks. Likewise, the use of

an electrode array distributed across the scalp allowed us to take advantage of spreading voltage distributions at the scalp in ERP measures, which is in fact often considered a disadvantage of the ERP technique. When a dipole is present in a conductive medium such as the brain, electricity does not just run directly between the two poles of a dipole, but instead spreads out through the conductor (Luck, 2005). ERPs spread out as they travel through the brain following the path of least resistance. Consequently, an ERP generated in one part of the brain can lead to substantial voltages at rather distant parts of the scalp (Luck, 2005), which could give rise to similar ERP responses measured at some locations even if they come from different sources. This is relevant to the so-called inverse problem. Instead of trying to deal with this problem, we took advantage of the spreading voltage at the scalp by including more electrodes in our analysis. It is important to note that we did not attempt to identify the locations and orientations of generator dipoles, as it is impossible to do so with the ERP technique. Instead, the idea behind this was that if the same neural mechanisms were recruited in syntactic processing of sentences in the AGL and NLSP tasks, the voltage distribution at the scalp elicited by the stimuli used in the two tasks should be the same.

Results from this extended analysis were consistent with the previous analysis using ERP responses obtained at Pz within one large 500-800 msec latency window. Neither the three-way interaction among task, scalp distribution (electrode sites), and conditions nor the two way interaction between task and scalp distribution reached significance. The statistically undistinguishable scalp distributions provided one piece of evidence suggesting that the same neural mechanism might have been recruited for both processing the natural language stimuli and the sentence strings in the AGL task.

6.2 Expanding in Time

The ERP technique provides a high temporal resolution of brain responses. Moving in time therefore means something very different from changing locations such

as expanding electrode sites discussed above to compare P600 effects between the AGL and the NLSP tasks. ERP responses captured at different times usually reflect different underlying processes. For instance, the ERP component N400 has a typical latency window of 300-500 msec after onset of words with semantic anomalies (Berkum, Hagoort, Brown, 1999; Osterhout and Holcomb, 1995). Although our focus on was on syntactic processing of sentences and therefore on the P600, it was also of interest to perform overall comparisons to see if there were differences between the two tasks on other ERP components. One component in particular, the (early) left anterior negativity (ELAN/LAN) effects, often observed over left anterior scalp sites, have been related to grammatical processing (Hahne & Friederic, 1999). The ELAN/LAN has been observed in association with phrase structure violation (Friederici et al., 1993, 1996; Münte, Heinze, & Mangun, 1993; Neville et al., 1991; Osterhout & Holcomb, 1992, 1993), with the processing of subcategorization information (Osterhout & Holcomb, 1993; Rösler, Friederici, Pütz, & Hahne, 1993), and with agreement violations (Coulson et al., 1998; Friederici et al., 1993;Gunter et al., 1997; Osterhout & Mobley, 1995). Critically, Friederici and colleagues have argued that P600 is a controlled process, while ELAN/LAN is specific to grammar. To this end, we used 4 windows with a latency of 300 msec (100-400 msec, 300-600 msec, 500-800 msec, 700-1000 msec) and included ERP responses obtained from a more spread-out scalp region, including left/central/right anterior, left/right central and central, and left/central/right posterior.

Using this expanded analysis we found task differences in only one time window. Specifically, we found early negativities within the 100-400 msec latency window for the NLSP task but not for the AGL task. The fact that this response was found only for the NLSP task suggests a task-specific effect. Since the words (*was*, *were*, *would*, *had*) that were time-locked to in the NLSP task for the ERP measure are all closed-classed words, this effect is reminiscent of a particular characteristic of the LAN effect elicited by processing of closed-class word that has been reported in the literature (Neville, Mills, &

Lawson, 1992; Nobre & McCarthy,1994). Given that the negativities in the NLSP task have a frontal scalp distribution and that it started relative early, it is possible that the observed negativities are reflective of ELAN/LAN effects. However, the negativities did not show a left anterior scalp distribution. Therefore, it is uncertain as to whether the observed negativities are ELAN/LAN effects.

6.3 The Corpus Analyses

Before we started the NLSP task, we conducted corpus analyses to confirm the categorizations of verbs in the study by Osterhout and colleagues. Different from their study in which verb categorizations were based on verb transitivity measures obtained in two sentence completion tasks, DO/SC probabilities (the probability of verb followed by an object noun versus a subject noun of a sentential complement clause) were calculated. A new verb grouping was formed based on results from the corpus analyses. One major change in the new grouping was that the Good Fit condition moved closer to the Excellent Fit condition. Interestingly, this change resulted in a corresponding change in ERP responses, with the ERP waveform of the Good Fit condition within the latency window of the P600 moving closer to that of the Excellent Fit condition. Although the P600 waveform moved in the expected direction, the magnitude differences of P600 between the Excellent Fit and Good Fit conditions became insignificant when the new grouping was applied.

There are two possible interpretations to this effect. First of all, the nonsignificant effect suggests that “verb transitivity” might better capture the state of the knowledge better than DO/SC categories. However, since the 3 corpuses together contained only 3 million words, the rather small sample size might not provide good enough basis to calculate DO/SC probabilities for the verbs.

A third, and perhaps a more sensible, interpretation is that the relationship between input statistics and neurophysiological responses is not linear. This is often seen

in work on computational modeling of neural networks in which communication between individual processing elements (neurons) are typically based on non-linear transfer functions that generates a single output value from all of the input values that are applied to the neuron. Nonlinearity is important because this allows sequences of neurons chained together to achieve more complex computations than single stages of neural processing can (O'Reilly & Munakata, 2000).

Taking a sigmoidal activation function for instance, when the input probability difference in the Excellent Fit and the Good Fit conditions was adjusted to be closer in the new grouping, the corresponding differences in neurophysiological responses might have become too subtle to show in ERP measures. After all, the P600 effect of the Good Fit condition in the new grouping did move in a predicted direction (toward the Excellent Fit condition), suggesting that the magnitude of P600 effects is sensitive to this change.

6.4 Implications for the Indexing Function of the P600 Component

The major ERP component we focused on in this thesis is the P600 component. In chapter 2, we summarized eliciting conditions of the P600 component, including a wide range of syntactic violations, including morphological violations of number, gender, and case (Atchley et al., 2006; Coulson et al., 1998; Osterhout and Mobley, 1995), phrase structure (Hagoort et al., 1993; Neville et al., 1991), subjacency (McKinnon & Osterhout 1996; Neville et al., 1991), and subcategorization (Ainsworth-Darnell et al., 1998; Osterhout & Holcomb, 1992). But what cognitive processes does the P600 reflect? There are at least two different proposals with regard to the indexing function of the P600 in the literature.

The first, and probably the most common view in the literature is that the P600 component associated with reanalysis processes. For instance, Friederici (1995) claimed that the P600 reflects repair processes following the detection of an ungrammatical

element. This claim is based on an observation that the P600 component often co-occurs with LAN (Coulson et al., 1998; Friederici et al., 1993; Münt, Heinze & Mangun, 1993). Since LAN occurs earlier in time than the P600 component, they argued that it is reasonable to associate the LAN with detection of ungrammaticality and the P600 with a repair process. Further support for this proposal came from a study by and Münt, Matzkem and Johannes (1997), who found a LAN, but no P600 for syntactic violations in German sentences in which all content words were replaced by nonce words but the inflectional morphology was kept intact. They argued that since all the content words were replaced with nonce words, it is likely that reanalysis only takes place if the words in the sentence actually make sense. As a result, the lack of P600 effect was taken as a piece of evidence that the P600 component reflects reanalysis processes. However, to our knowledge, there are at least two studies using sentences generated by an artificial grammar that found a P600 effect (Christiansen, Conway, & Onnis, 2007; Friederici, Steinhauer, & Pfeifer, 2002). Although the two studies both involved visual presentations of graphic symbols, we did not provide our participants with any visual referent corresponding to the words or the sentence string in our AGL task. Still, P600 effects were seen in the AGL task used in this thesis. In addition, the proposed reprocessing does not provide a feasible interpretation for magnitude variations of the P600. A second interpretation of the P600 is that it reflects the cost of reprocessing (Osterhout et al., 1994). Compared with the first proposal, one advantage of a resource-based account is that it not only captures the major cognitive process (reprocessing), but also provides a feasible account for magnitude differences of the P600 component. However, if reprocessing only takes place if sentences actually carry meaning, we should not have obtained the P600 effects in our AGL task.

In this thesis, we argued that the P600 associates with the well-documented expectation generation process in processing materials involving structural regularities. This proposal is based on convergent evidence reported in previous studies showing that

the P600 is sensitive to the amount of experience one has with a particular kind of materials and can be elicited by nonlinguistic materials. Within this context, we argued that the magnitude of the P600 varies as a function of degree of deviation between a target item and an expected one(s) based on the preceding material in the utterance. We referred to this as "goodness of fit" throughout this thesis. Although we did not directly test this assumption, the results obtained from the two major tasks in this thesis have some value in revealing the sources that give rise to magnitude variations of the P600: statistical learning of structural regularities. Even if we assume that reprocessing always takes place even in processing sentences without a transparent meaning, the results in the current study suggest that the amount of resources devoted to perform reprocessing is regulated by what is learned and represented. Our best guess on what happened during on-line sentence processing is that there is moment by moment expectation generation for the subsequent items based on already unfolded materials. The magnitude of the P600 effects therefore might reflect the amount of resources devoted to adjust the expected activation to match what is actually realized. This proposal requires further evaluation in future studies before a more certain conclusion can be drawn.

6.5 Limitations and Future Directions

In the NLSP task, two different verb groupings were used. When the original grouping was used, we obtained graded P600 responses across conditions. Different from the 1994 study by Osterhout and colleagues, we obtained a significant difference between the Excellent Fit and Good Fit conditions. However, when the new grouping was applied, the magnitude differences of P600 between the Excellent Fit and Good Fit conditions became insignificant. As discussed earlier, one possible explanation for the disappearance of the effect is that that "verb transitivity", as compared to DO/SC probabilities, better reflects the state of the knowledge. However, this account requires further examination because the difference between the Excellent Fit (intransitive) and the Good Fit

(intransitively-biased) conditions was not significant when “verb transitivity” was applied in the 1994 study by Osterhout and colleagues. It is also noted that the 3 corpora used in the current study together were composed of only 3 million words and therefore might not constitute a representative enough sample to evaluate statistical distributions of the verbs. Future studies should further evaluate this interpretation by including copora of a larger sample size. A third explanation is that the relationship between input statistics and neurophysiological responses is not linear. In this thesis, we did not have data to further evaluate this interpretation. One possible future direction is to use stimuli with statistical distributions sufficiently spread out and varied to capture fine-grained changes of the relationship between the input statistics and brain responses along the probability continuum.

In this thesis, we went beyond the P600 effect and compared overall ERPs elicited in the AGL and the NLSP tasks. This was done by including two additional time frames, 100-400 msec and 300-600 msec, in the comparisons. We found a main effect of task, contributed by a more negative mean voltage in the NLSP task than the AGL task within the 100-400 latency window. As discussed earlier, the early negativities might reflect the fact that the words that ERPs were time-locked to in the NLSP task are closed-class words. This could be treated as evidence for symbolic representations of word categories in processing natural language, which is absent in the artificial grammar. However, the negativities did not show a left anterior scalp distribution. Therefore, it is uncertain as to whether the observed negativities are the same ELAN/LAN effects reported in the literature. Even the early negativities truly reflect ELAN/LAN effects, the extent to which one could argue for symbolic representations based on our results is still a question.
A closed class usually contains a relatively *small number* of items. It is not impossible that this characteristic of a “closed-class” could be detected and learned based on input statistics (of a different kind). In fact, Monaghan, Chater, and Christiansen (2005) have

shown that phonological cues and distributional information are helpful for detecting grammatical categories.

Another factor that might have contributed to between task differences in ERPs is a timing difference. Taking a processing perspective, the magnitude of the P600 was considered in this thesis as a reflection of the goodness of fit (or the degree of violation of expectation) of a subsequent target item (i.e., a target word) based on the statistics (of the preceding materials) accumulated to the point right before hearing the target item. Consider the below sentences:

(a) The traveler remembered _____[$_{\text{zero, noun, WH- clause, adv…}}$]
(b) The traveler remembered the house___ [$_{\text{zero/ BE}}$].

The matrix verb "remembered" is a strong expectation generator of the subsequent structure such as a noun phrase (e.g., *The traveler remembered <u>the house</u>*) and a *wh*- clause (e.g., The traveler remembered *<u>which terminal to go</u>*). The type of structural competition focused on whether or not there was a subsequent copula verb (such as "*was*") following the postverbal noun phrase (*the house*). In this study, we included a corpus analysis to calculate the probability of occurrence of a sentential complement clause (e.g., *The traveler remembered the house <u>was old</u>*) after a "matrix verb + noun" constituent. These statistics were different from those computed by Osterhout and colleagues in their 1994 study where the probability of occurrence of a sentential complement clause after the matrix verb alone (i.e., Transitive/Intransitive probabilities) was used. The fundamental difference between these two calculations lies in whether or not the occurrence of the postverbal noun phrase (the house) was taken into consideration. We argued that "matrix verb + noun" constituents provided a better basis than matrix verb alone to capture the competition between the DO and the SC structure. Theoretically, the competition between DO and SC structures was not 'settled" until the

occurrence of a post-verbal noun phrase. The occurrence of a postverbal noun phrase would give rise to a more focused competition between DO/SC structures as other structures such as *wh*-clause would have been eliminated from the competition with the occurrence of a postverbal noun phrase. Therefore, we argued that the use of DO/SC probabilities could better reflect expectation generation processes during sentence processing than the use of verb transitivity. One might argue that right after the occurrence of the matrix verb "*remember*", the prediction process might not just stop at the next item immediately followed the matrix verb, but would keep going down the stream. If this is the case, we would expect to see a timing difference between the AGL task and the NLSP task because in the AGL task there was no intervening item between the word carrying a strong predictive value (C1, C2, E1, E2) and the word that the ERPs were time-locked to, while in the NLSP task there was an intervening noun phrase. The results reported in 4.42 where the P600 effects elicited in the AGL and the NLSP tasks were compared using smaller latency windows (500-600, 600-700, 700-800 msecs) provided evidence in support of our argument. There was no evidence of between task differences in terms of the P600 effects within different latency windows.

Finally, although the current study provided some evidence supporting the statistical account of grammatical acquisition, it is not clear at which level(s) the statistics operate on for generating expectation. For the AGL task, the manipulation of the probabilities among experimental conditions was based on lexical levels: between C1/C2 or E1/E2 and the subsequent word. For the NLSP task, however, the level of probability differences among the 4 experimental conditions was less clear. It is certain that the 4 conditions in the NLSP task differed with regard to the matrix verbs included in each condition. However, the expectation generation process could operate on abstract categories (*Copula BE*) or specific lexical items (*was*), or, perhaps more likely, both in parallel. In other words, the predictive relationship could be lexically-based, categorical-based, or between lexical and category. This question is fundamental to understanding the

nature of language (or grammatical) representations. Future studies are required in order to further explore this question.

APPENDIX A
TABLES

Table A1. Categorization of syntactic frames.

Category	Subcategorization frames	Example
Direct object (DO)	VP (specifier) NP	She couldn't *believe* her eyes.
	VP NP NP	They offered to *buy* him a warehouse
	VP NP PP	I *followed* its progress from the chase car.
	NP modifier	They were called upon to *decide* the issues immediately.
	VP NP that-S	Jenny Brice *knew* the necklace that was passed down to my mother.
	Perception complement	I didn't see them leave.
Sentential complement	VP (that) SC	The bankers also *insist* that the probability is low. She admitted she made a big mistake but *insisted* her motives were correct
other	VP NP Infin-S	He *forced* me to take the job
	VP wh-s	Pretoria hasn't *forgotten* why they were all sentenced to life imprisonment in the first place.
	Nominal	The economy itself seems locked in a struggle between *hope* and fear.
	Passive	Some panic will be *seen*.
	VP if…	They won't *buy* it if the price is too high.
	NP front	Some of the coal the companies *buy* will come from Westmoreland mines.
	Verb phrase	I never *heard of* it.
	0	He *agreed*.
	Adjectival participle	That is the *promised* land.
	Quote	Promoters *believe*: "Ernest Ball was."
	idiom / frozen phrases	We don't *see* eye to eye to each other
	Question	Do you *remember* him?
	Supplementation	That's the reason, I *think.*
	Discourse fillers	Getting there, you *know*, is quite difficult.
	incomplete/revision/truncated	He discussed…we discussed that with him.
	insertion/inversion	He is happy, I *guess*, to receive that.
	different meaning	electrical *charge*; The committee had a *hearing* last week;

Table A2. Results of corpus analyses: Percent DO and intransitive usage.

Osterhout et al. (1994) verb categories		SWBD			WSJ		
categories	verbs	DO	Intranstive	Transitivity bias	DO	Intranstive	Transitivity bias
Transitive	Agree	21.43	78.57		0.00	100.00	
	decide	6.11	93.89		7.19	92.81	
	hope	0.56	99.44		0.00	100.00	
	insist	0.00	100.00		0.00	100.00	
	think	1.28	98.72	94.12 (9.02)	0.69	99.31	98.42 (3.15)
Intransitively-biased	believe	38.03	61.97		3.49	96.51	
	guess	5.99	94.01		8.33	91.67	
	know	37.55	62.45		42.76	57.24	
	promise	66.67	33.33		38.98	61.02	
	remember	61.81	38.19	57.99 (24.15)	82.14	17.86	64.85(31.63)
Transitively-biased	charge	100.00	0.00		100.00	0.00	
	forget	70.00	30.00		61.11	38.89	
	hear	72.68	27.32		90.74	9.26	
	see	95.67	4.33		95.97	4.03	
	understand	58.59	41.41	79.38 (17.71)	66.67	33.33	82.89(17.76)
Transitive	buy	100.00	0.00		100.00	0.00	
	discuss	100.00	0.00		100.00	0.00	
	follow	100.00	0.00		100.00	0.00	
	force	100.00	0.00		99.02	0.98	
	include	100.00	0.00	100 (0)	100.00	0.00	99.80 (0.43)

Osterhout et al. (1994) verb categories		BC			All 3 corpora		
categories	verbs	DO	Intranstive	Transitivity bias	DO	Intranstive	Transitivity bias
Transitive	Agree	3.70	96.30		3.84	96.16	
	decide	7.35	92.65		6.93	93.07	
	hope	1.77	98.23		0.66	99.34	
	insist	0.00	100.00		0.00	100.00	
	think	12.66	87.34	94.90(5.03)	2.15	97.85	97.28 (2.78)
Intransitively-biased	believe	18.27	81.73		21.05	78.95	
	guess	10.00	90.00		6.13	93.87	
	know	48.02	51.98		40.15	59.85	
	promise	39.58	60.42		40.71	59.29	
	remember	56.03	43.97	65.61(19.59)	61.30	38.70	66.13 (21.04)
Transitively-biased	charge	61.54	38.46		96.27	3.73	
	forget	66.67	33.33		67.78	32.22	
	hear	88.89	11.11		81.53	18.47	
	see	85.95	14.05		93.83	6.17	
	understand	82.46	17.54	77.10(12.21)	69.26	30.74	81.73 (13.30)
Transitive	buy	99.07	0.93		99.95	0.05	
	discuss	100.00	0.00		100.00	0.00	
	follow	99.15	0.85		99.68	0.32	
	force	93.62	6.38		97.81	2.19	
	include	100.00	0.00	98.36(2.69)	99.85	0.15	99.45 (0.92)

Table A3. Percent DO/SC probabilities for the twenty verbs used in Osterhout et al. (1994) study.

Category	Verb exemplar	DO %	SC %	Mean *SC* Probability (*SD*)
Intransitive	agree	9.72	90.28	94.46 (6.45)
	decide	14.83	85.17	
	hope	1.01	98.99	
	insist	0	100	
	think	2.16	97.84	
Intransitively-biased	believe	21.12	78.88	57.65 (29.72)
	guess	6.14	93.86	
	know	40.37	59.63	
	promise	73.02	26.98	
	remember	71.12	28.88	
Transitively-biased	charge	96.27	3.73	16.12 (11.30)
	forget	78.21	21.79	
	hear	81.53	18.47	
	see	94.15	5.85	
	understand	69.26	30.74	
Transitive	buy	100	0	0.18 (0.26)
	discuss	100	0	
	follow	99.68	0.32	
	force	99.44	0.56	
	include	100	0	

Note: Individuals verb were organized into four categories based on the same categorization reported in the original study.

Table A4. Results of verb re-subcategorization based on the results from the corpus analyses.

Category	Verb exemplar	DO %	SC %	Mean SC Probability (*SD*)
Intransitive	agree	9.72	90.28	96.19 (4.04)
	hope	1.01	98.99	
	insist	0	100	
	think	2.16	97.84	
	*guess	6.14	93.86	
Intransitively-biased	believe	21.12	78.88	70.03 (8.73)
	know	40.37	59.63	
	*suspect	23.47	76.53	
	*doubt	26.79	73.21	
	*admit	38.19	61.90	
Transitively-biased	* promise	73.02	26.98	25.37 (5.11)
	*remember	71.12	28.88	
	forget	78.21	21.79	
	hear	81.53	18.47	
	understand	69.26	30.75	
Transitive	buy	100	0	0.18 (0.26)
	discuss	100	0	
	follow	99.68	0.33	
	force	99.44	0.56	
	include	100	0	

Note: Verbs with an asterisk are new verbs or verbs from a different category as in Osterhout et al. (1994) study

Table A5. Percent mean accuracy and standard deviations of the grammaticality judgments in the AGL and NLSP Tasks.

		Excellent	Good	Low	Bad	*Total*
		% Mean Accuracy (SD)				
AGL	Pre-training Test	47.31 (9.14)	49.42 (10.13)	48.55 (8.96)	52.77 (4.92)	49.51 (3.83)
	Post-training Test	94.78 (3.36)	96.75 (2.53)	95.97 (4.04)	95.00 (3.01)	95.93 (2.55)
NLSP	Original grouping	95.86 (2.47)	93.26 (4.34)	93.26 (2.81)	94.72 (5.97)	94.28 (2.57)
	New grouping	95.18 (2.58)	93.69 (3.74)	95.49 (2.63)	94.70 (3.40)	94.77 (1.50)

Table A6. Mean amplitude in uV at Pz within 500-800 msec latency window in the pre- and post-training tests of the AGL task.

	Excellent	Good	Low	Bad
	Mean Amplitude (uV)			
Pre-training Test	.67 (2.44)	-.08 (2.54)	-.28 (2.63)	.40 (1.92)
Post-training Test	.52 (.92)	1.69 (1.26)	2.54 (1.02)	3.56 (1.18)

Table A7. Mean amplitude in uV at Pz within 500-800 msec latency window in the NLSP task.

	Excellent	Good	Low	Bad
	Mean Amplitude (uV)			
Original grouping	-.15 (.82)	1.58 (1.48)	2.52(1.35)	4.09 (1.54)
New grouping	.14 (1.51)	.68 (1.43)	2.86 (1.09)	4.09 (1.54)*

* No verbs were replaced in the Bad Fit condition from the original grouping to the new grouping

Table A8. Difference waves for conditions Good Fit minus Excellent Fit, Low Fit minus Excellent Fit, and Bad Fit minus Excellent Fit for the AGL (post-training) and NLSP (original grouping) tasks.

		Good-Excellent	Low-Excellent	Bad-Excellent
		Mean Amplitude (uV)		
AGL	Pre-training Test	1.17(1.21)	2.01 (1.21)	3.04 (1.47)
NLSP	Original grouping	1.73 (1.70)	2.67 (1.64)	4.24 (1.85)

Table A9. Mean amplitudes and standard deviations for the Good-Excellent (G-E), Low-Excellent (L-E), and Bad-Excellent (B-E) conditions at C3, CZ, C4, P3, PZ, P4, O1, OZ, and O2 within 500-600, 600-700, and 700-800 msec latency windows in the AGL task.

AGL	C3	Cz	C4	P3	Pz	P4	O1	Oz	O2
	Mean amplitude (SD) in uV								
500-600 (msec)									
G-E	1.02 (1.17)	.72 (.94)	.65 (1.02)	.57 (1.77)	1.30 (1.39)	1.10 (1.46)	.57 (1.26)	.76 (1 40)	1.25 (1.78)
L-E	1.01 (1.96)	1.15 (1.32)	.97 (1.53)	1.69 (1.49)	1.99 (1.20)	1.92 (1.43)	1.65 (1.43)	1.71 (1.22)	1.71 (.97)
B-E	2.07 (1.84)	2.14 (1.74)	1.86 (1.68)	2.63 (1.42)	3.12 (1.72)	2.74 (1.87)	1.38 (1.17)	2.03 (1.08)	2.31 (1.81)
600-700 (msec)									
G-E	1.23 (1.17)	1.29 (1.27)	1.01 (1.38)	.67 (1.62)	1.63 (.96)	1.13 (.96)	.66 (1.02)	.79 (.98)	1.19 (1.63)
L-E	1.43 (1.60)	2.09 (1.25)	1.75 (1.29)	2.10 (1.52)	2.47 (1.33)	2.27 (1.11)	1.76 (1.70)	1.80 (1.05)	1.80 (1.06)
B-E	2.36 (1.57)	2.89 (2.09)	2.48 (1.90)	3.29 (1.34)	4.30 (1.64)	3.57 (1.59)	1.78 (1.32)	2.49 (1.40)	2.91 (2.00)
700-800 (msec)									
G-E	1.04 (1.06)	1.13 (.94)	1.12 (1.55)	.43 (1.58)	1.53 (.96)	1.13 (.93)	.43 (1.54)	.61 (1.07)	1.19 (1.73)
L-E	1.11 (1.72)	1.87 (1.40)	1.67 (1.60)	1.79 (1.45)	2.34 (.96)	2.25 (1.12)	1.52 (1.63)	1.66 (.97)	1.81 (.91)
B-E	2.22 (1.86)	2.72 (1.96)	2.41 (1.940	3.24 (1.71)	4.17 (1.58)	3.64 (1.50)	1.75 (1.46)	2.48 (1.38)	2.99 (1.85)

Table A10. Mean amplitudes and standard deviations for the Good-Excellent (G-E), Low-Excellent (L-E), and Bad-Excellent (B-E) conditions at C3, CZ, C4, P3, PZ, P4, O1, OZ, and O2 within 500-600, 600-700, and 700-800 msec latency windows in the NLSP task.

NLSP	C3	Cz	C4	P3	Pz	P4	O1	Oz	O2
	Mean amplitude (SD) in uV								
500-600 (msec)									
G-E	1.35 (1.64)	1.35 (1.74)	.75 (1.73)	1.34 (2.08)	1.77 (1.84)	1.17 (1.43)	.89 (1.76)	1.30 (1.52)	.63 (1.44)
L-E	2.20 (1.73)	1.88 (1.82)	1.55 (1.07)	2.60 (2.24)	3.16 (2.10)	2.46 (1.55)	1.33 (2.35)	1.80 (1.66)	1.28 (2.09)
B-E	2.64 (1.86)	3.18 (2.09)	2.57 (1.77)	3.23 (1.38)	4.45 (1.72)	3.33 (1.51)	1.71 (1.51)	2.04 (1.58)	1.88 (1.83)
600-700 (msec)									
G-E	1.12 (1.62)	1.15 (1.72)	.59 (1.62)	1.03 (2.20)	1.64 (1.86)	.94 (1.35)	.70 (1.60)	1.22 (1.41)	.49 (1.25)
L-E	2.26 (1.54)	1.77 (1.46)	1.60 (1.03)	2.32 (1.96)	2.97 (1.68)	2.34 (1.31)	1.10 (2.42)	1.80 (1.45)	1.27 (1.86)
B-E	2.75 (2.25)	3.22 (2.52)	2.68 (2.33)	3.21 (1.56)	4.86 (2.16)	3.70 (1.79)	1.71 (1.72)	2.36 (1.68)	2.13 (1.72)
700-800 (msec)									
G-E	1.14 (1.41)	1.39 (1.60)	.70 (1.87)	1.14 (1.89)	1.83 (1.80)	.93 (1.41)	.46 (1.38)	1.14 (1.44)	.42 (1.18)
L-E	1.53 (1.19)	1.28 (1.45)	.95 (1.38)	1.51 (1.83)	2.13 (1.73)	1.58 (1.30)	.27 (2.38)	1.19 (1.34)	.82 (1.93)
B-E	3.10 (2.34)	3.47 (2.77)	2.61 (2.820	3.44 (1.48)	5.10 (2.23)	3.70 (2.14)	1.74 (1.72)	2.53 (1.55)	2.35 (1.47)

Table A11. A summary of results from some previous studies using an AGL task.

	Saffran (2001b)	Dienes, Broadbent, & Berry (1991)	Friederici , Steinhauer, & Pfeifer (2002)	Christiansen, Conway, & Onnis (2007)	Whittlesea & Dorken (1993)	Gomez , Gerken, & Schvaneveldt (2000)	Brooks & Vokey (1991)	Vokey & Higham (2005)
Stimuli type	Spoken nonword sentences	Visual display of letter strings	Spoken nonword sentences and visual referents	Visual display of nonword sentence with visual scenes	Visual display of letter strings	Visual display of letter strings	Visual display of letter strings	Visual display of letter strings
Experimental conditions	Experiment 1; intentional condition	Experiment 1; (Grammatical group)	n/a	n/a	Experiment 2	Experiment 2 (strings containing no versus one repeating element)	n/a	n/a
Training time	2 sessions, 30 mins each	Not specified	Several training sessions (up to 5 hours each) until a 95% accuracy criterion was reached	30 mins	Not specified	Not specified	Self-paced	Self-paced
Behavioral measure	Pair string acceptability judgment	Grammaticality Judgments	Grammaticality Judgments	Grammaticality Judgments	Grammatical-ity Judgments	Grammatical-ity Judgments	Grammatical-ity Judgments	Grammatical-ity Judgments
Accuracy (%)	66.67-75%	60%	93%	93.9%	Correct acceptance:64 %; Incorrect acceptance :27%	Correct acceptance: 55-73%; Incorrect acceptance :26-53%	Correct acceptance:48 -61%; Incorrect acceptance:33 -46%	Correct acceptance: 49-61%; Incorrect acceptance:35 -52%

APPENDIX B
FIGURES

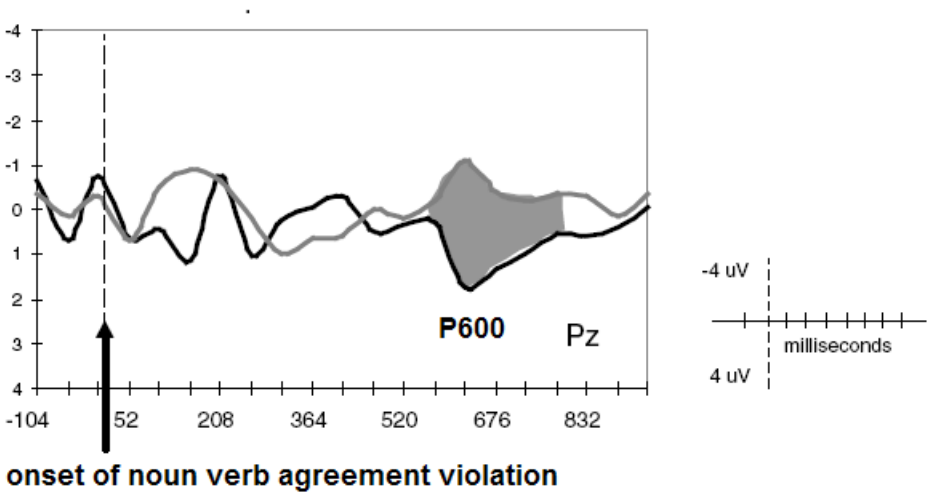
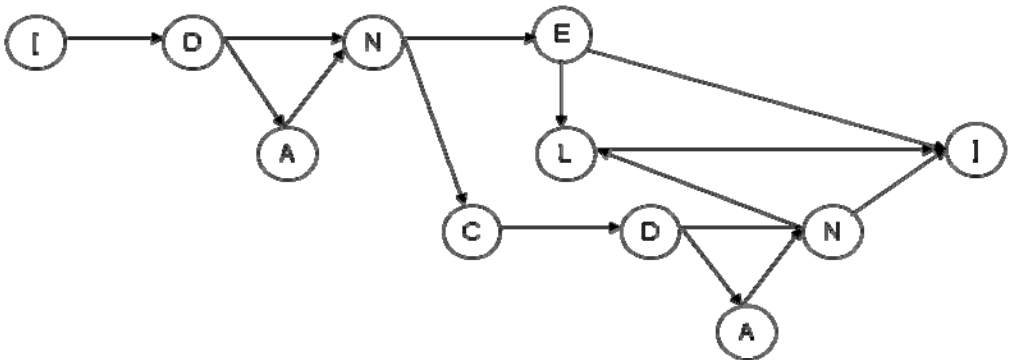


Figure B1. A sample figure of ERPs to noun-verb agreement violation. The gray line indicated averaged ERPs of well-formed sentences (e.g., *She returns the book*) and the black line indicated averaged ERPs of sentences containing agreement violation (e.g., *She return the book*). The shaped area represented time period during which the two waveforms were significantly diverging.



Vocabularies:
D = {tok, osh}
A ={ak, vat}
N ={glif, jos, tish, pel}
C_1, C_2 = {plox, baf}
E_1, E_2 ={zitch, trul}
L = {tood, rix, lum}

Figure B2. The state diagram of the artificial grammar used in this study. Single letters represent different nodes. These nodes specify word classes and the arrows indicate valid transitions between nodes. Every sentence starts from the beginning node ([) to the end node (]).

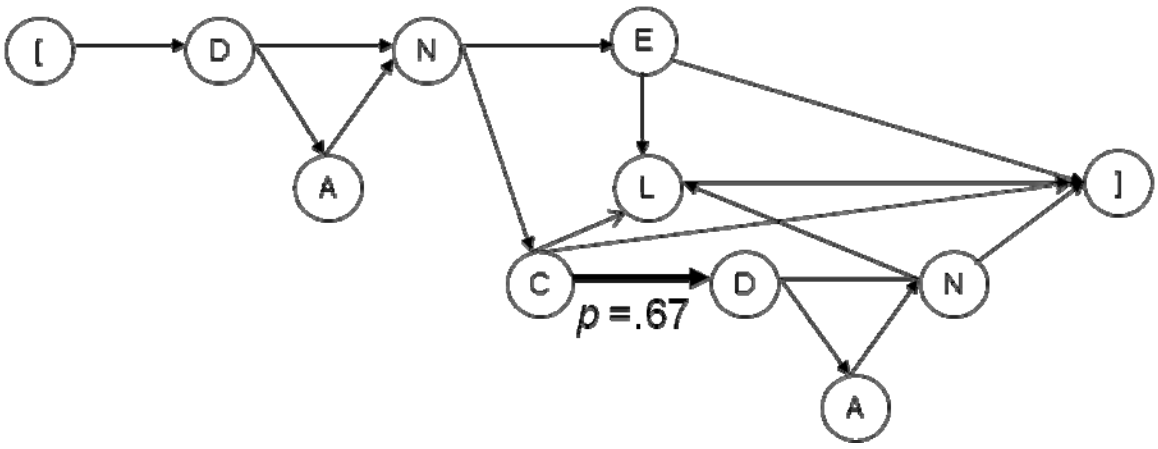


Figure B3. The state diagram with probabilistic patterns specified for $C_{2.}$ The bold arrow indicates alternative route that is invalid for the prototypical word member C1 but is valid for $C_{2.}$

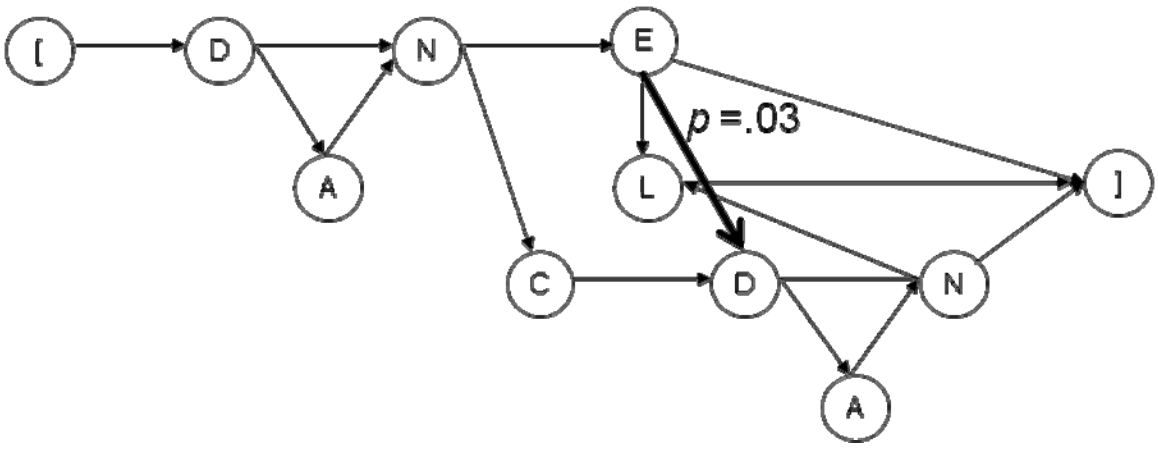


Figure B4. The state diagram with probabilistic patterns specified for $E_{2.}$ The bold arrow indicates alternative route that is invalid for the prototypical word member E_1 but is valid for $E_{2.}$

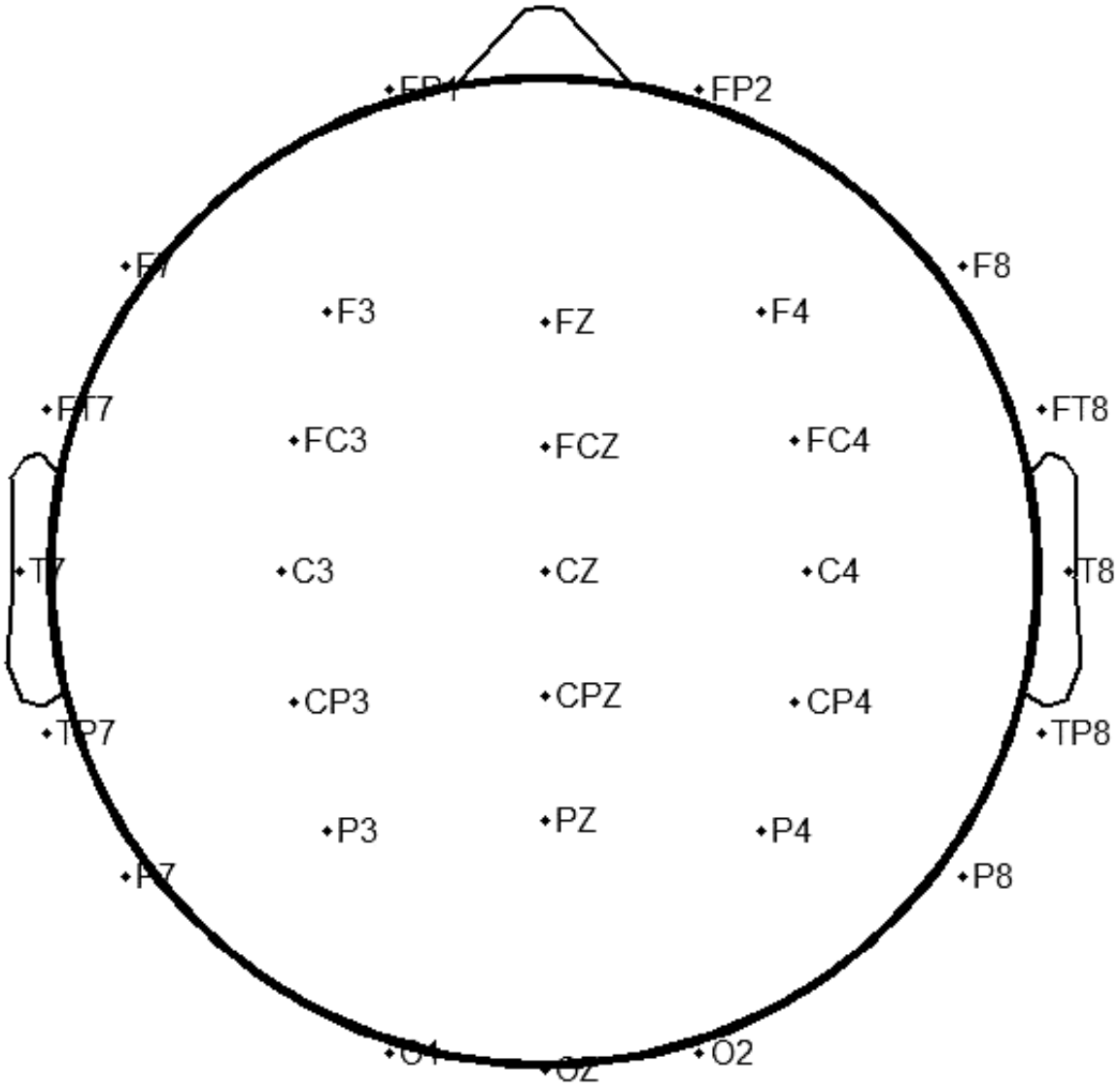


Figure B5. Schematic diagram of electrode montage used in this study.

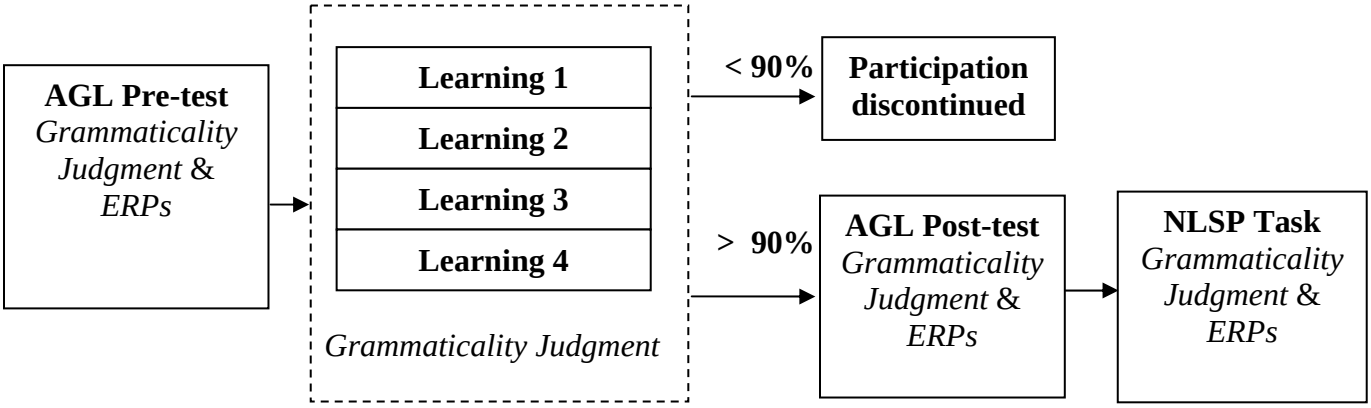


Figure B6. Overall procedure of the AGL and NLSP tasks. Words in italic indicates measures given in each phase.

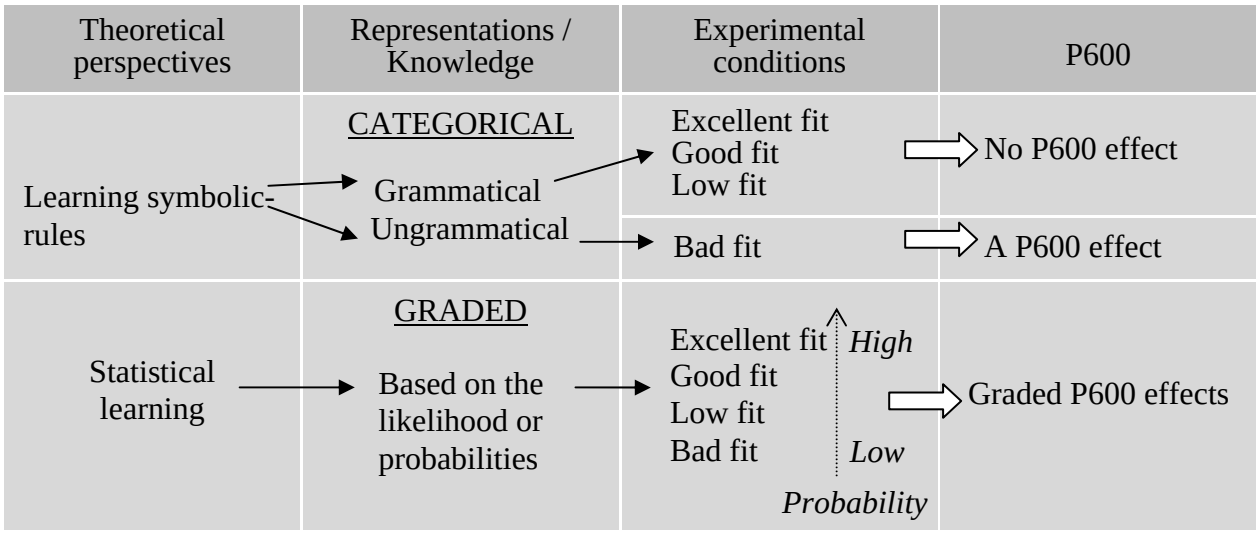


Figure B7. Predictions of P600 effects in terms of the two contrastive theoretical accounts of grammatical acquisition.

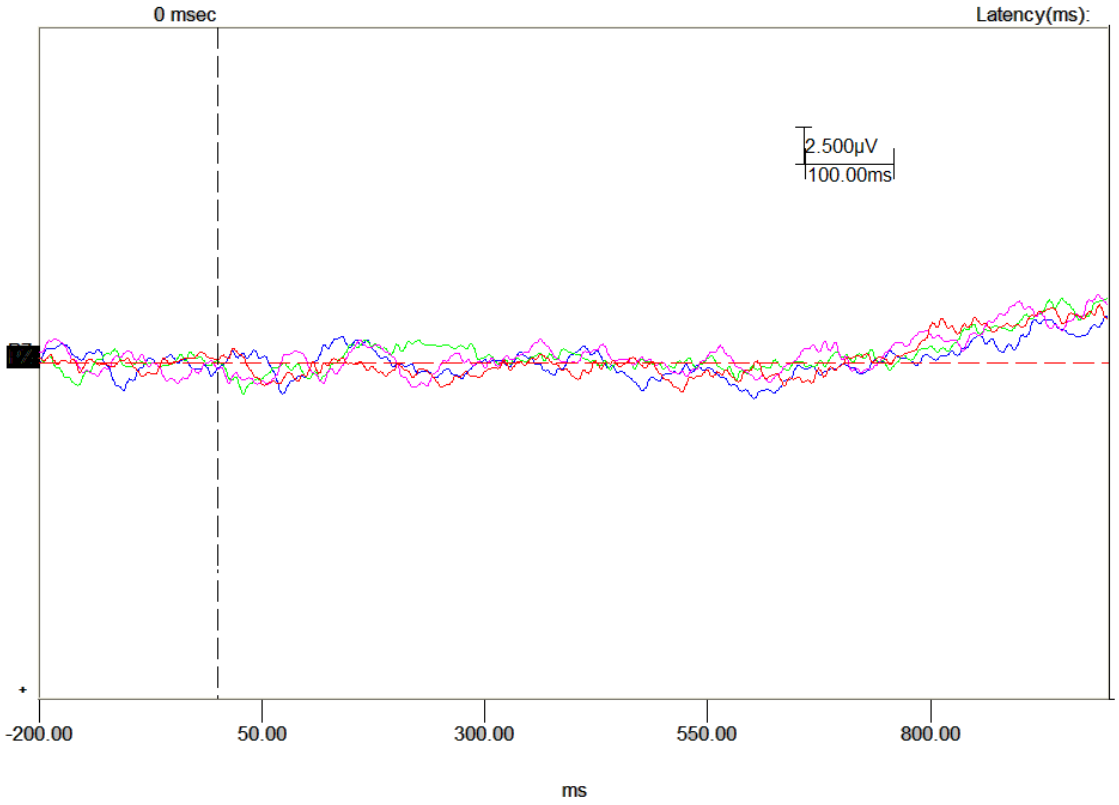


Figure B8. ERPs in the pre-training test of the AGL task. Grande average event-related brain potentials (across all participants and items) from the parietal P_Z site for the target word in the Excellent Fit (baseline), Good Fit, Low Fit, and Bad Fit conditions.

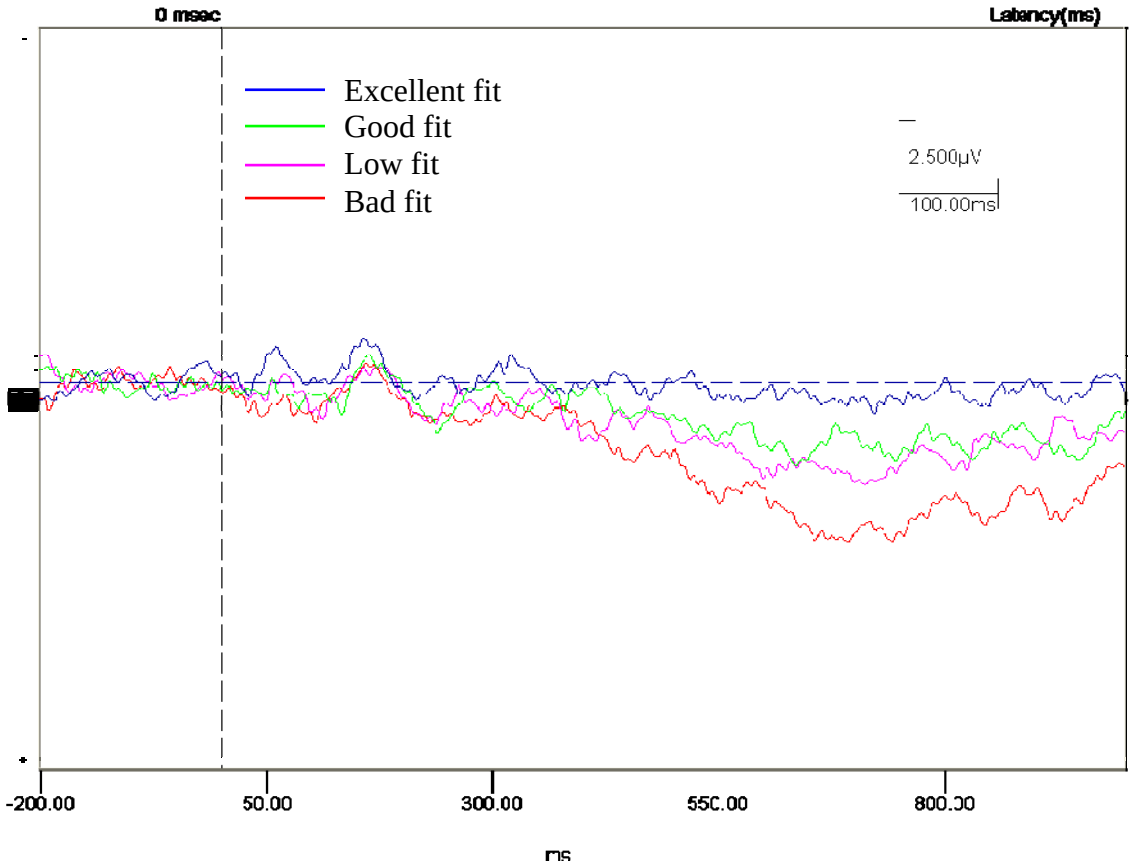


Figure B9. ERPs in the post-training test of the AGL task. Grande average event-related brain potentials (across all participants and items) from the parietal P_Z site for the target word in the Excellent Fit (baseline), Good Fit, Low Fit, and Bad Fit conditions.

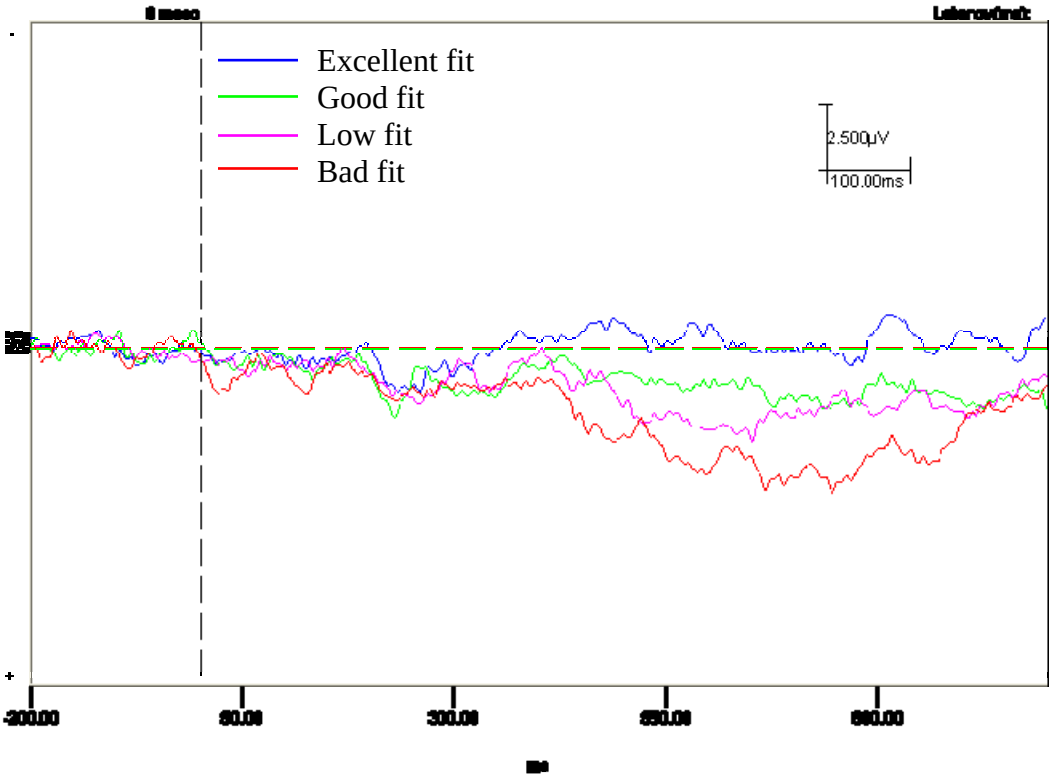


Figure B10. ERPs in the NLSP task using the original verb grouping. Grande average event-related brain potentials(across all participants and items) from the parietal P_Z site for the target word in the Excellent Fit (baseline), Good Fit, Low Fit, and Bad Fit conditions.

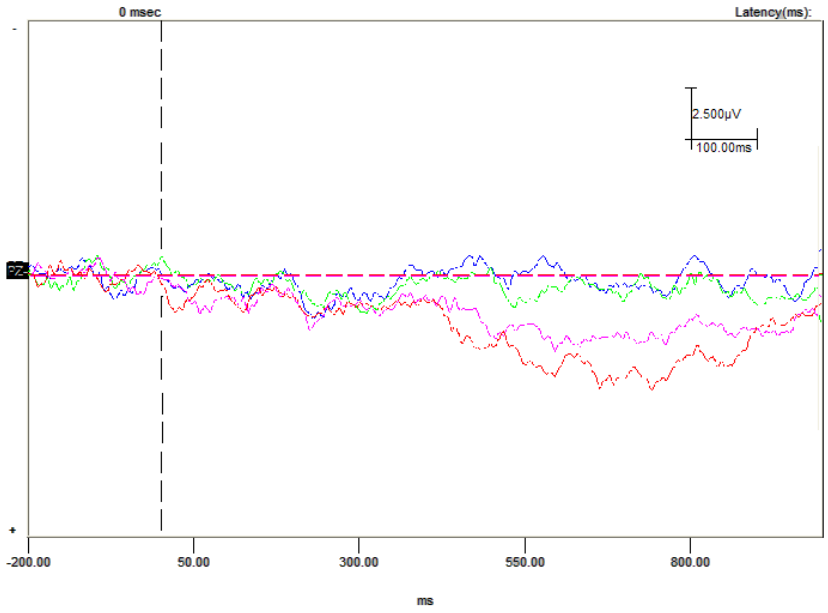


Figure B11. ERPs in the NLSP task using the new verb grouping. Grand average event-related brain potentials (across all participants and items) from the parietal P_Z site in the Excellent Fit (baseline), Good Fit, Low Fit, and Bad Fit conditions in the NLSP task.

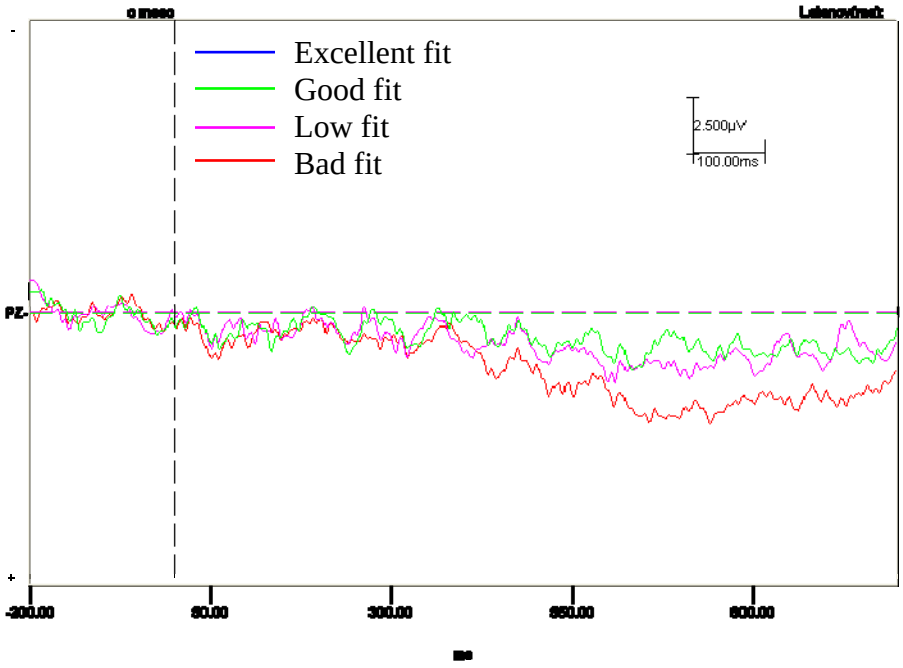


Figure B12. ERPs based on difference waves in the AGL task (post-training). Difference waves from the parietal Pz site were formed by subtracting ERPs elicited by the target word in the Excellent condition from ERPs elicited by target word in the Good fit (green line), Low fit (pink line), and Bad fit (red line) condition.

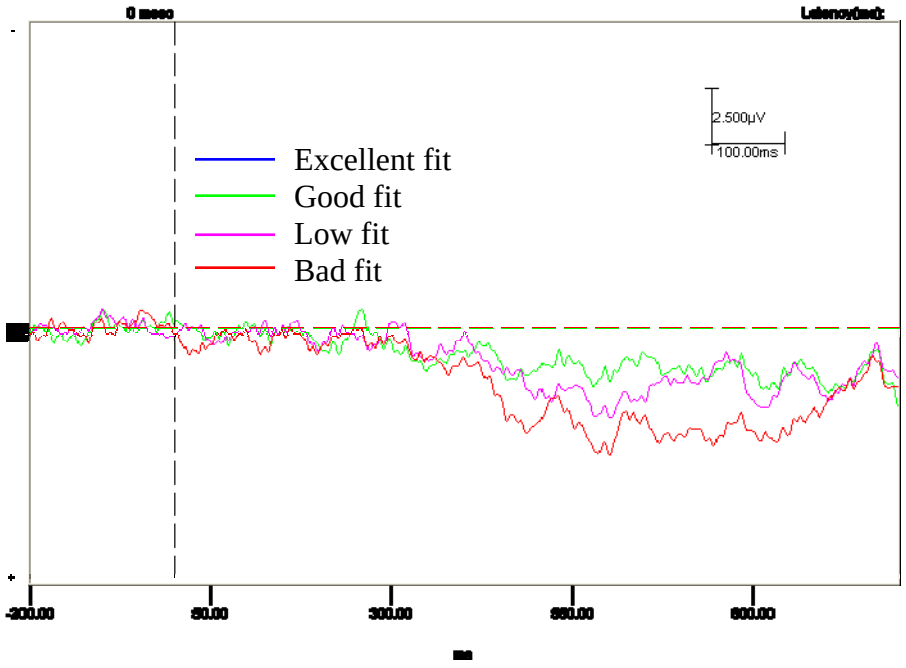


Figure B13. ERPs based on difference waves in the NLSP task (original grouping). Difference waves from the parietal Pz site were formed by subtracting ERPs elicited by the target word in the Excellent condition from ERPs elicited by target word in the Good fit (green line), Low fit (pink line), and Bad fit (red line) condition in the NLSPTask.

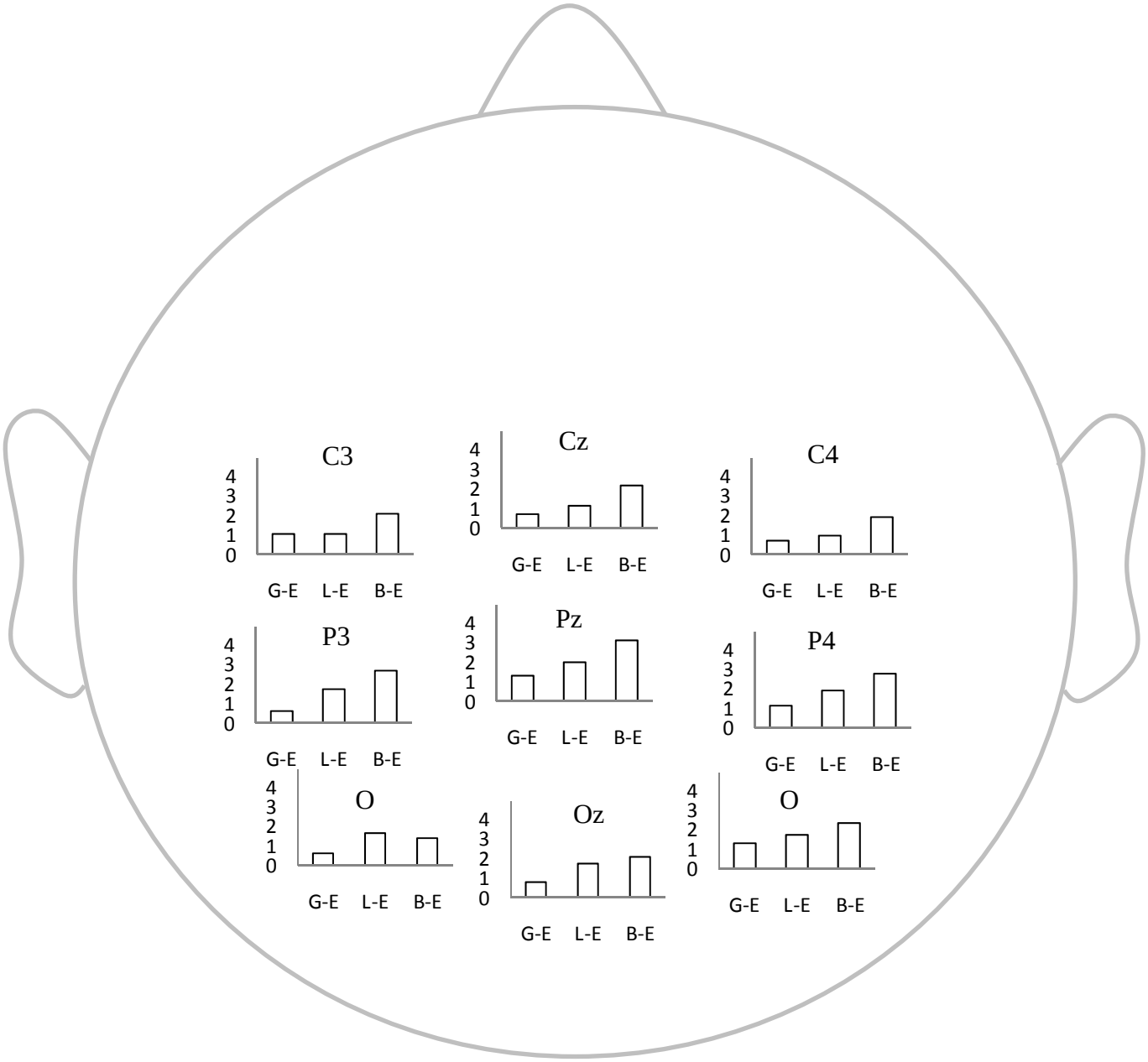


Figure B14. Difference waves for the Good-Excellent (G-E), Low-Excellent (L-E), and Bad-Excellent (B-E) conditions at C3, CZ, C4, P3, PZ, P4, O1, OZ, and O2 in the AGL task within 500-600 msec window.

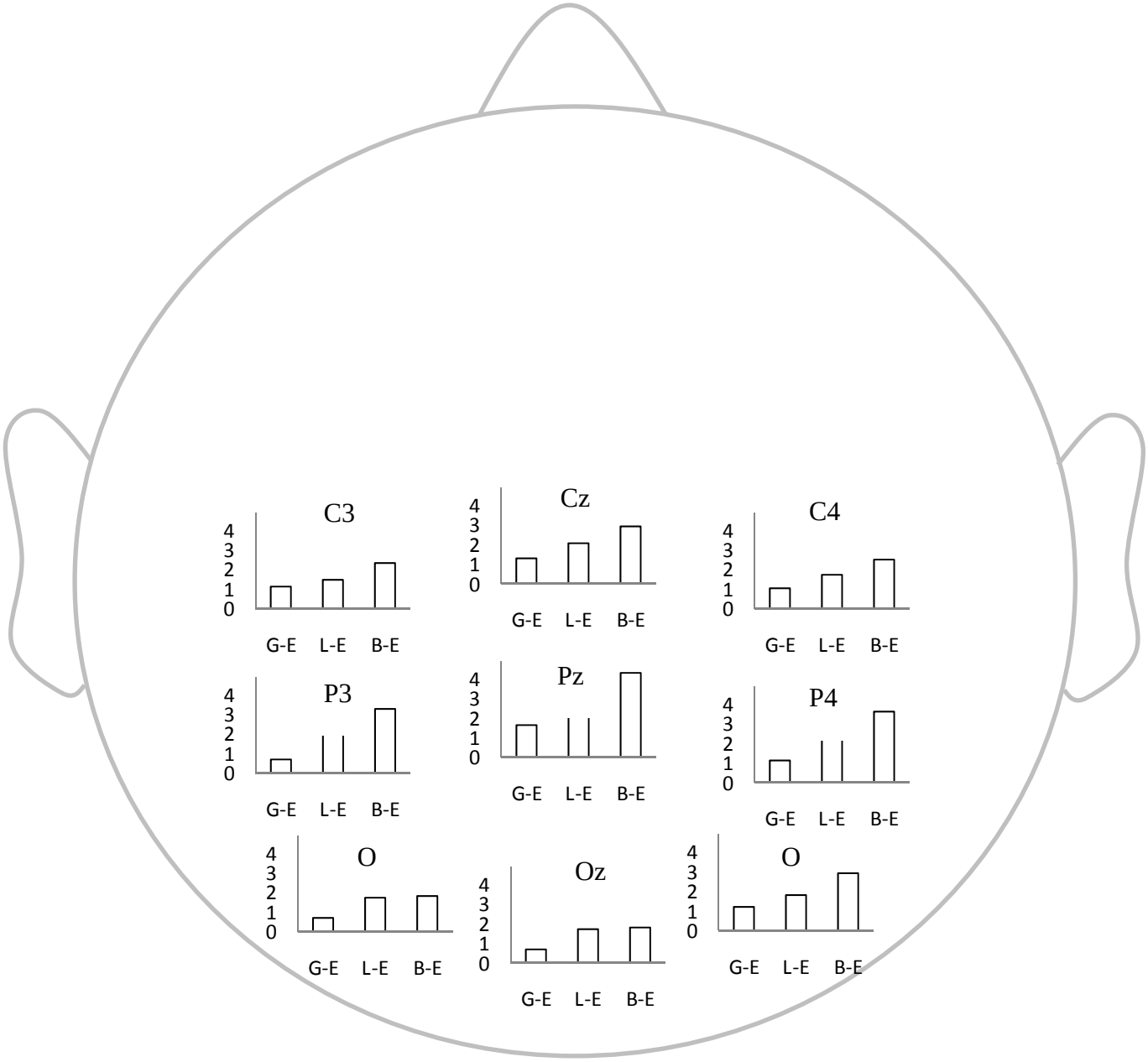


Figure B15. Difference waves for the Good-Excellent (G-E), Low-Excellent (L-E), and Bad-Excellent (B-E) conditions at C3, CZ, C4, P3, PZ, P4, O1, OZ, and O2 in the AGL task within 600-700 msec window.

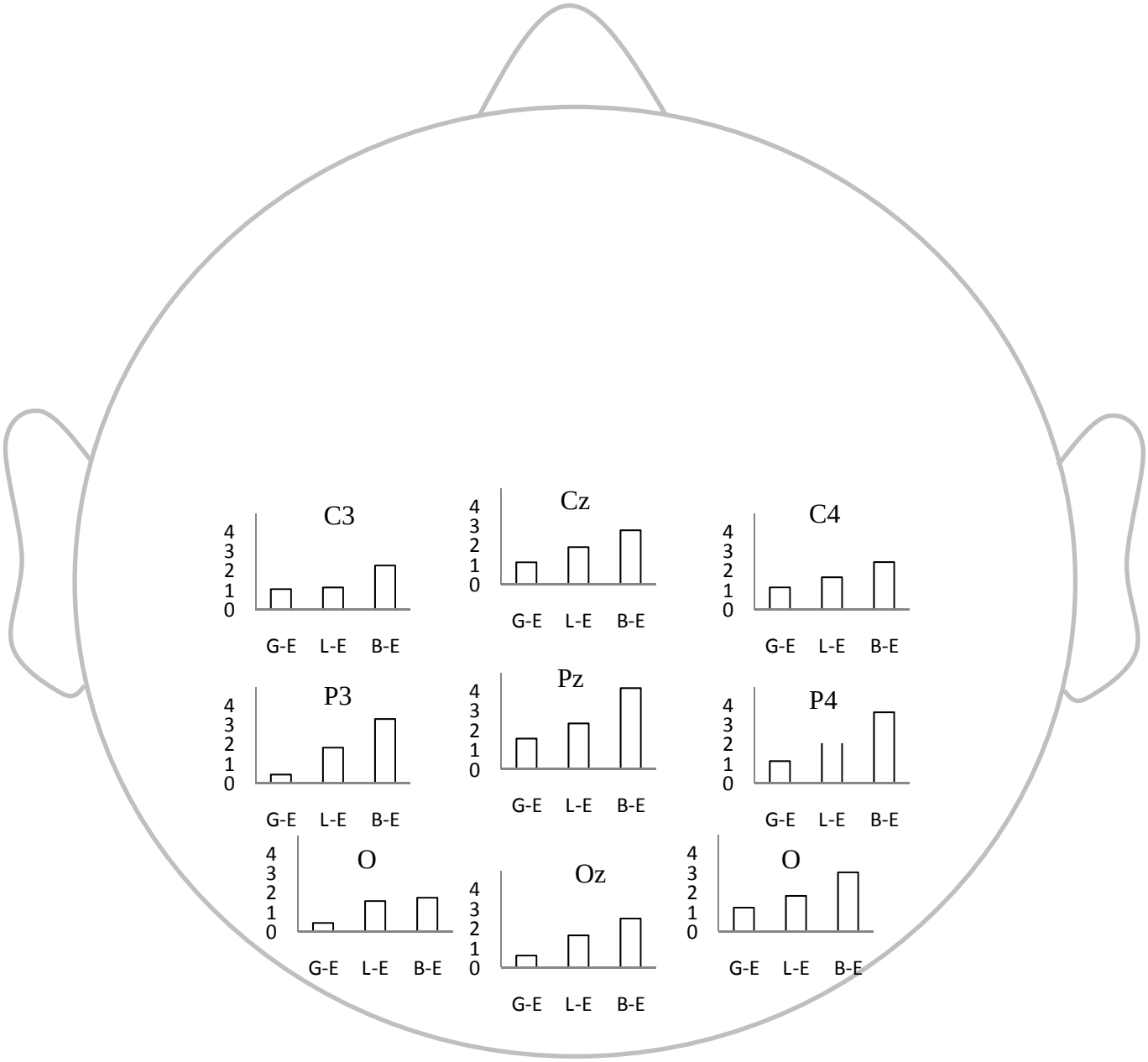


Figure B16. Difference waves for the Good-Excellent (G-E), Low-Excellent (L-E), and Bad-Excellent (B-E) conditions at C3, CZ, C4, P3, PZ, P4, O1, OZ, and O2 in the AGL task within 700-800 msec window.

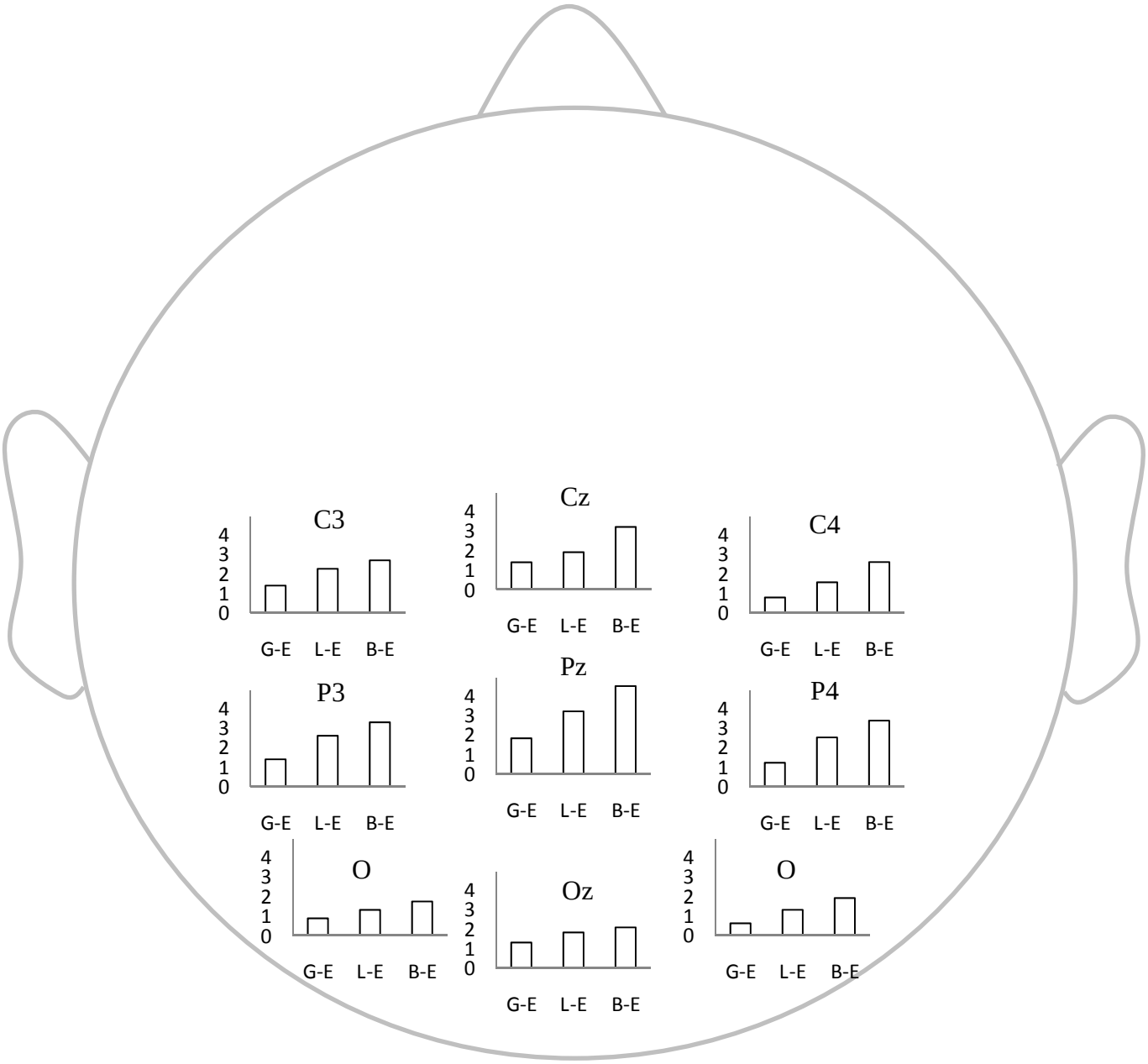


Figure B17. Difference waves for the Good-Excellent (G-E), Low-Excellent (L-E), and Bad-Excellent (B-E) conditions at C3, CZ, C4, P3, PZ, P4, O1, OZ, and O2 in the NLSP task within 500-600 msec window.

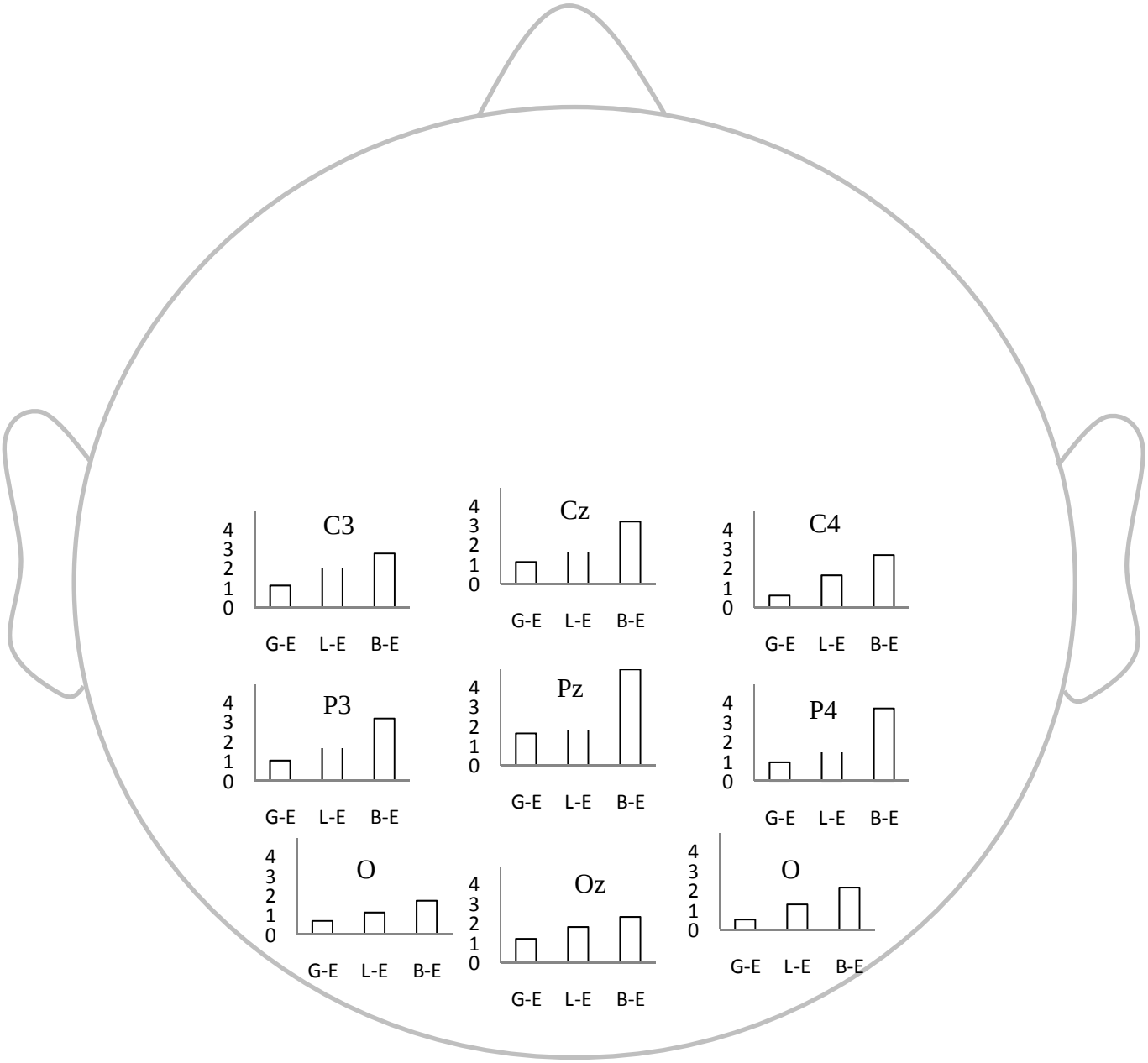


Figure B18. Difference waves for the Good-Excellent (G-E), Low-Excellent (L-E), and Bad-Excellent (B-E) conditions at C3, CZ, C4, P3, PZ, P4, O1, OZ, and O2 in the NLSP task within 600-700 msec window.

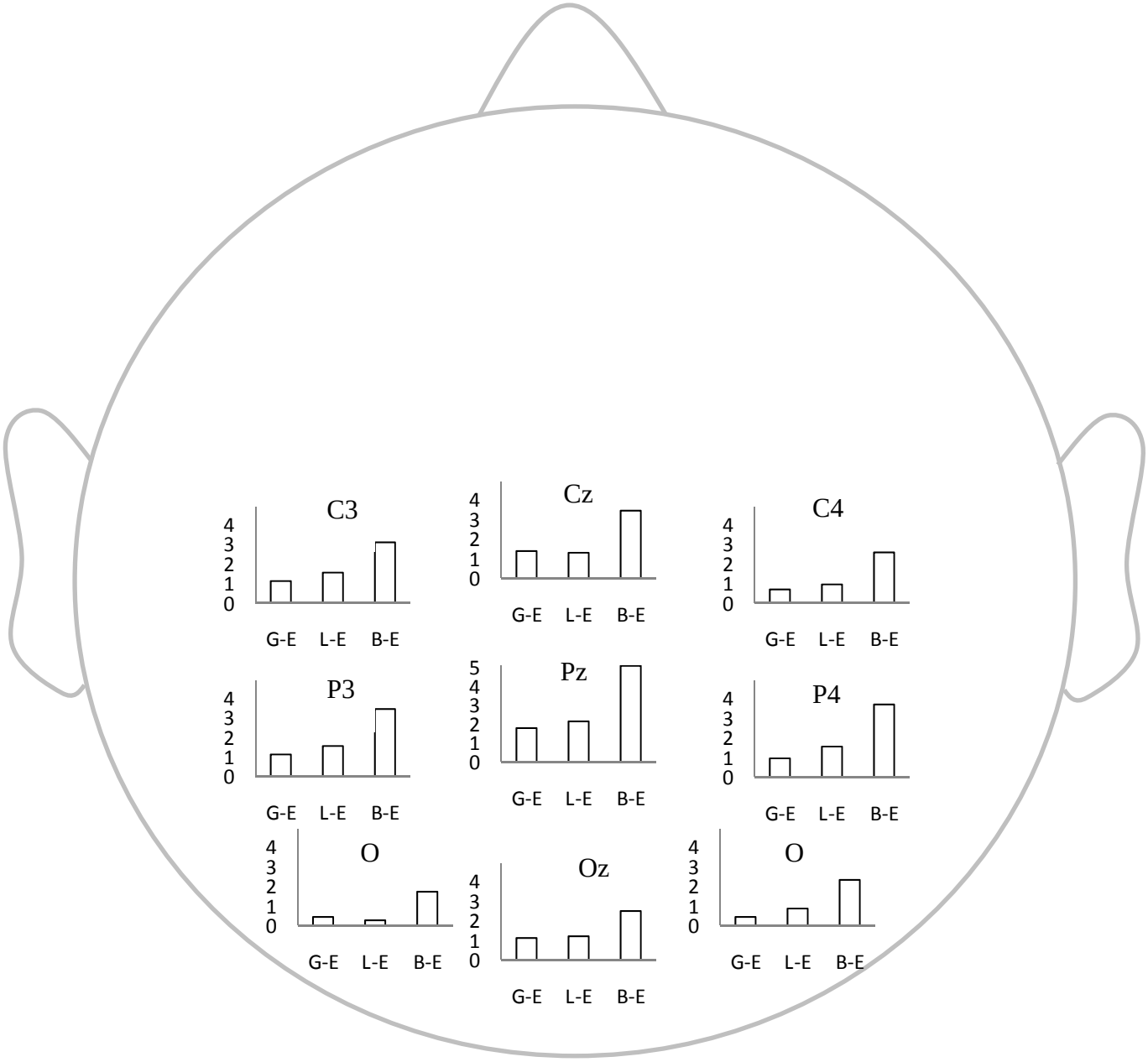


Figure B19. Difference waves for the Good-Excellent (G-E), Low-Excellent (L-E), and Bad-Excellent (B-E) conditions at C3, CZ, C4, P3, PZ, P4, O1, OZ, and O2 in the NLSP Task within 700-800 msec window.

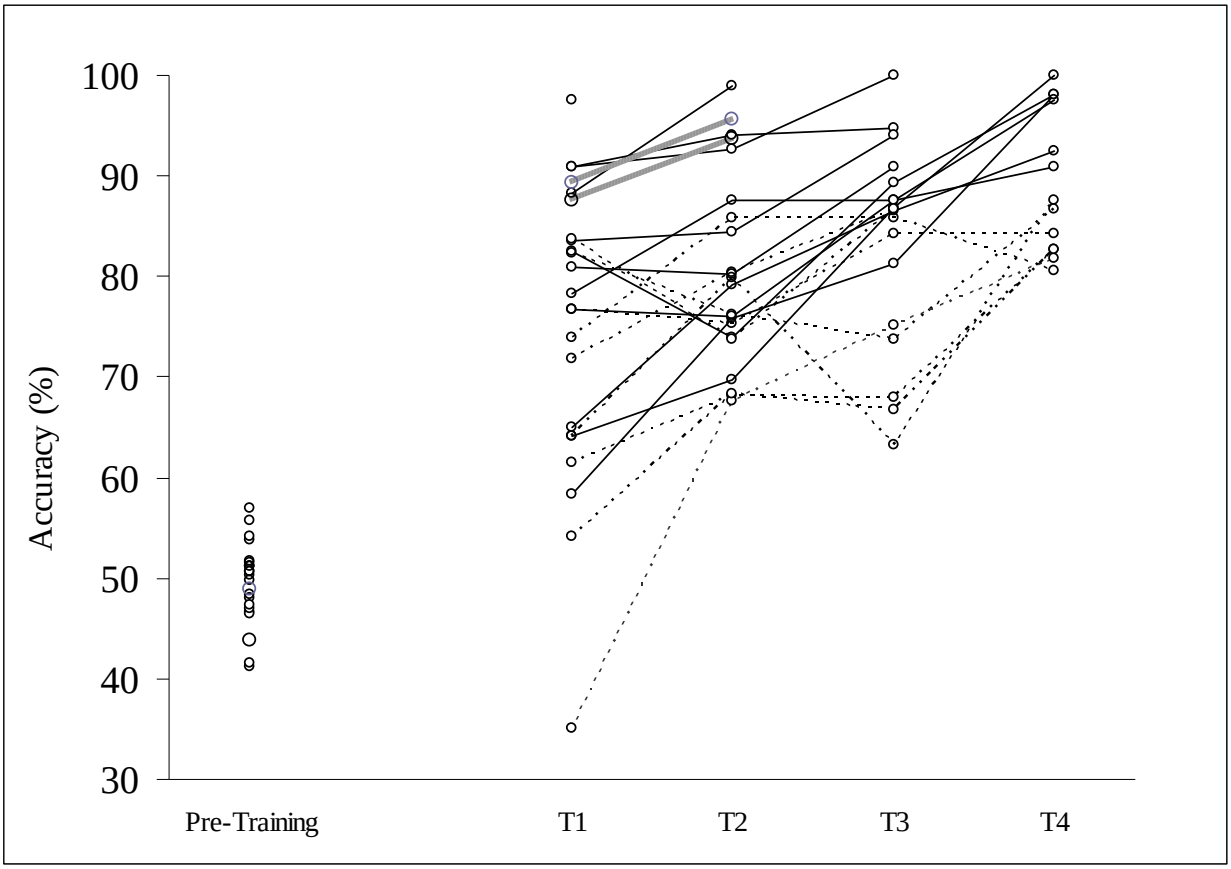
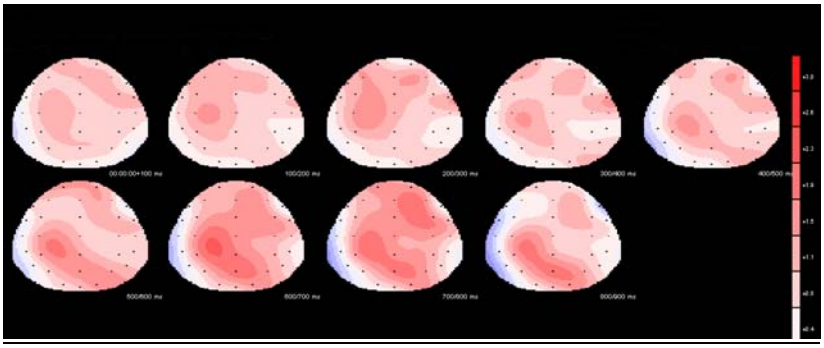
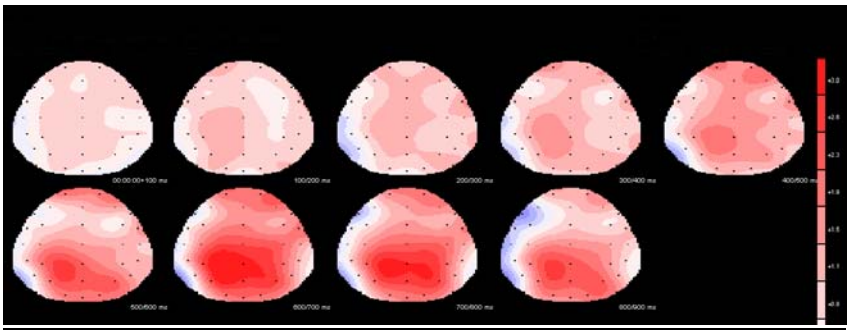


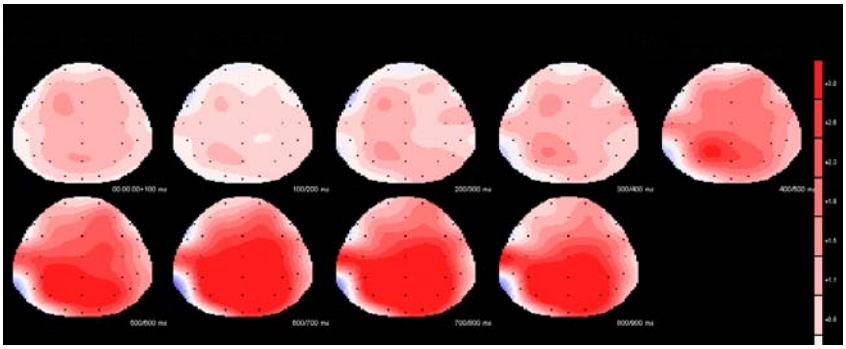
Figure B20. Proportions correct for individual participants across training sessions.



(a) AGL: Good Fit-Excellent Fit

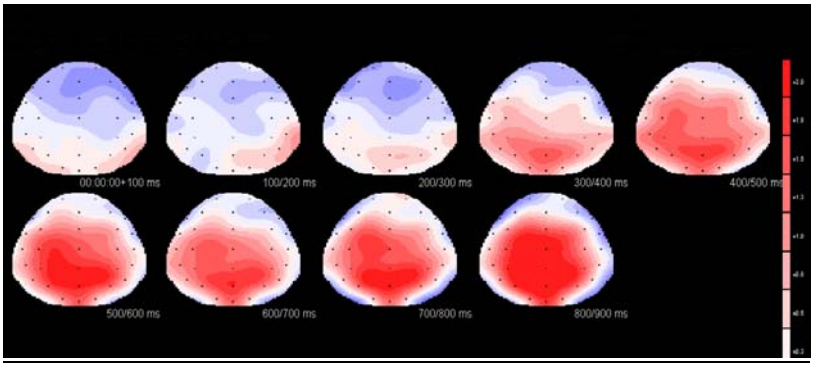


(b) AGL: Low Fit-Excellent Fit

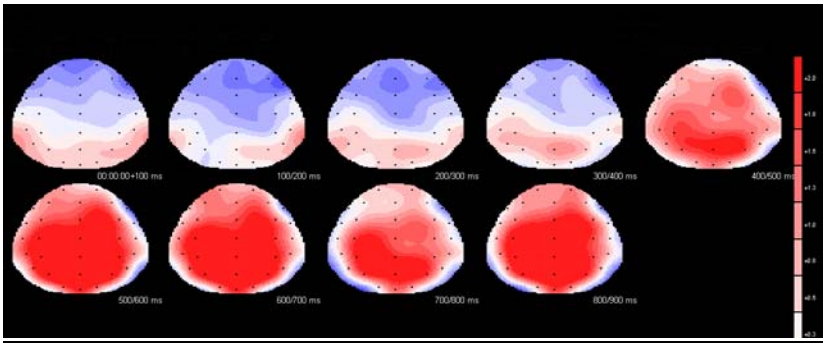


(c) AGL: Bad Fit-Excellent Fit

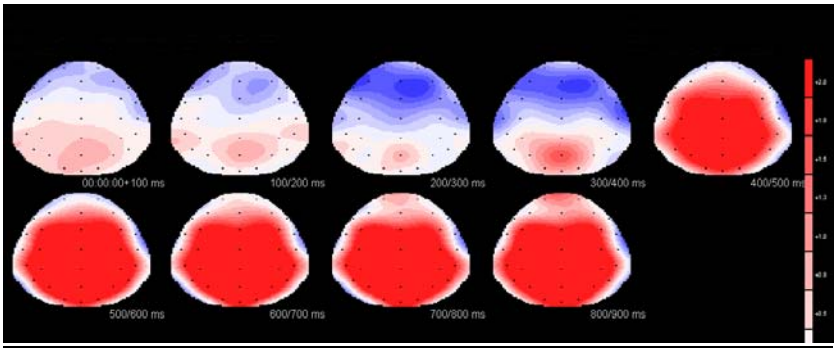
Figure B21. Topography of the scalp distribution for the target word with a 100 msec step from 0 msec to 1000 msec for the (a) Good Fit, (b) Low Fit and (c) Bad Fit conditions after subtracting from the Excellent Fit condition in the AGL task.



(a) NL: Good fit-Excellent fit



(b) NL: Low fit-Excellent fit



(c) NL: Bad fit-Excellent fit

Figure B22. Topography of the scalp distribution for the target word with a 100 msec step from 0 msec to 1000 msec for the (a) Good Fit, (b) Low Fit, and (c) Bad Fit conditions after subtracting from the Excellent Fit condition in the NLSP task.

APPENDIX C

TARGET SENTENCES USED IN THE NLSP TASK

No.	Sentences
1	The captain agreed/bought/believed/heard the crew was unhappy.
2	The banker hoped/discussed/knew/forgot the secretary had called.
3	The musician thought/followed/remembered/heard the sonata was beautiful.
4	The mother insisted/included/promised/saw the child would sleep.
5	The student decided/forced/guessed/understood the answer was incorrect.
6	The detective thought/bought/knew/charged the criminal was lying.
7	The physicist insisted/discussed/knew/understood the theory was wrong.
8	The doctor decided/followed/remembered/forgot the prescription had changed.
9	The man hoped/included/promised/forgot his wife would return.
10	The plumber agreed/forgot/guessed/saw the faucet was fixed.
11	The shopper insisted/bought/remembered/charged the bicycle was broken.
12	The psychiatrist decided/discussed/believed/saw the patient was sane.
13	The landlord thought/followed/knew/heard the tenant was angry.
14	The company hoped/included/promised/forgot the customer would arrive.
15	The detective agreed/forced/guessed/charged the criminal was guilty.
16	The judge agreed/followed/believed/understood the defendant was guilty.
17	The professor thought/discussed/knew/understood his lecture was unprepared.
18	The waitress agreed/bought/remembered/forgot the coffee was bitter.
19	The salesman insisted/included/promised/saw the child was safe.
20	The secretary decided/forgot/guessed/heard the number had changed.
21	The reporter decided/bought/believed/heard the story was inaccurate.
22	The quarterback hoped/discussed/knew/saw the receiver was open.
23	The traveler agreed/followed/remembered/understood the country was poor.
24	The grocer thought /included/believed/forgot the customer had shoplifted.
25	The spy insisted/forced/promised/understood his government was willing.
26	The woman decided/bought/guessed/charged the song was worthless.
27	The conductor hoped/discussed/knew/heard the orchestra was good.
28	The boy thought/followed/remembered/saw the dog was hungry.
29	The man hoped/forced/promised/forgot the police would arrive.
30	The accountant insisted/included/guessed/charged the purchases were unneeded.
31	The doctor hoped/discussed/believed/charged the patient was lying.
32	The workers insisted/forced/believed/charged the employer was sincere.
33	The scientist decided/included/knew/heard several solutions were possible.
34	The woman agreed/bought/remembered/forgot the key was lost.
35	The woman thought/included/knew/forgot the profits had disappeared.
36	The judge decided/bought/believed/understood the manuscript was misleading.
37	The professor insisted/followed/remembered/charged the student had failed.
38	The student agreed/discussed/knew/heard the assignment was insufficient.
39	The traveler hoped/bought/promised/saw the doctor was safe.
40	The soldier agreed/followed/knew/understood the decision was unfair.
41	The doctor agreed/followed/guessed/heard the patient was insane.
42	The salesman insisted/bought/promised/charged the car was safe.
43	The critic thought/discussed/guessed/understood the book was good.
44	The mechanic insisted/forced/promised/saw the car was fixed.
45	The banker thought/followed/guessed/understood the market would collapse.

46	The lawyer decided/forced/remembered/charged the defendant was guilty.
47	The customer thought/included/remembered/forgot the coffee was bitter.
48	The writer hoped/discussed/promised/heard the captain would succeed.
49	The secretary decided/included/guessed/understood the joke was inappropriate.
50	The plumber insisted/forced/believed/charged the housewife was lying.
51	The thief insisted/forced/knew/saw the combination had changed.
52	The host hoped/included/knew/forgot the guests would arrive.
53	The coach thought/discussed/believed/saw the professor was unfair.
54	The senator agreed/discussed/promised/heard the bill would pass.
55	The man agreed/included/guessed/understood the fee was excessive.
56	The actor hoped/followed/remembered/forgot the girl was married.
57	The doctor decided/forced/remembered/charged the patient was healthy.
58	The editor decided/bought/believed/forgot the manuscript was lost.
59	The workers insisted/bought/promised/saw the child had changed.
60	The officer insisted/followed/guessed/charged the dog was stolen.
61	The auditors insisted/discussed/believed/understood the deficit had increased.
62	The police hoped/forced/knew/charged the criminal would return.
63	The agent thought/followed/promised/heard the actress would arrive.
64	The lawyer agreed/included/guessed/charged the claims were false.
65	The pilot agreed/discussed/remembered/forgot the weather had worsened.
66	The actress thought/discussed/knew/heard the law was unfair.
67	The girl hoped/forced/remembered/saw the bank had closed.
68	The captain agreed/bought/promised/saw the child was safe.
69	The shopper decided/bought/knew/charged the merchandise was stolen.
70	The secretary decided/followed/remembered/understood the visitor was anxious.
71	The student agreed/forced/believed/heard the decision was wrong.
72	The investor hoped/bought/remembered/forgot the newspaper had failed.
73	The pilot decided/included/guessed/forgot the flight was delayed.
74	The doctor hoped/included/promised/saw the boy would heal.
75	The senator insisted/forced/believed/charged the secretary was lying.
76	The consultant agreed/decided/knew/heard the contract was inadequate.
77	The general insisted/discussed/guessed/understood the decision was risky.
78	The officer decided/followed/believed/charged the driver was drunk.
79	The waitress thought/forced/knew/saw the customers would leave.
80	The student insisted/discussed/believed/heard the war was unfair.
81	The mailman thought/followed/guessed/forgot the dog was unleashed.
82	The employer hoped/forgot/promised/understood the workers would improve.
83	The professor agreed/included/guessed/saw the student would succeed.
84	The sailor thought/bought/remembered/saw the ship was leaving.
85	The gambler insisted/included/promised/charged the game was fixed.
86	The cook decided/bought/promised/heard the recipe was good.
87	The professor decided/forced/knew/saw the student had cheated.
88	The housewife thought/bought/remembered/forgot her watch was broken.
89	The judge hoped/included/knew/understood the charges were dropped.
90	The banker insisted/followed/believed/forgot the accountant was stealing.
91	The governor agreed/discussed/knew/understood the problem was severe.

92	The man thought/bought/remembered/forgot his wallet was empty.
93	The farmer hoped/discussed/knew/saw the corn had grown.
94	The lawyer insisted/followed/believed/charged the senator was stealing.
95	The student hoped/included/guessed/forgot the concert was free.
96	The president insisted/forced/promised/understood the country would survive.
97	The woman insisted/followed/believed/heard the song was sexist.
98	The student agreed/bought/guessed/forgot the book was overdue.
99	The king thought/included/promised/saw the queen would complain.
100	The coach decided/forced/promised/understood the team would lose.
101	The activists insisted/discussed/believe/charged the mayor was unjust.
102	The nurse thought/forced/knew/saw the patient had improved.
103	The vacationer agreed/bought/remembered/forgot the hotel was old.
104	The soldier hoped/followed/promised/heard the general would return.
105	The actress decided/included/guessed/heard the song would succeed.
106	The general thought/forced/believed/charged the enemy was weak.
107	The baron insisted/discussed/knew/charged the countess had lied.
108	The librarian thought/included/remembered/saw the books were unshelved.
109	The sheriff decided/followed/guessed/understood the hermit was homeless.
110	The president hoped/discussed/promised/heard the shareholders were willing.
111	The cook agreed/bought/promised/forgot the boy was ready.
112	The student decided/included/guessed/forgot the answer was unknown.
113	The operator agreed/forced/knew/understood the caller was correct.
114	The reporter hoped/bought/believed/saw the story was big.
115	The prosecution insisted/forced/believed/charged the witness had lied.
116	The playboy thought/followed/remembered/forgot the woman was penniless.
117	The driver insisted/discussed/remembered/saw the accident was unavoidable.
118	The nurse hoped/followed/promised/heard the surgeon would improve.
119	The executive agreed/bought/guessed/charged the decision was poor.
120	The mayor agreed/included/knew/understood the situation was serious.

REFERENCES

Ainsworth-Darnell, K., Shulman, H. G., & Boland, J. E. (1998). Dissociating brain responses to syntactic and semantic anomalies: Evidence from event related potentials. *Journal of Memory and Language, 38*, 112-130.

Altmann G (2002a). Learning and development in neural networks—the importance of prior experience. *Cognition, 85*, B43-B50.

Altmann, G. T. M. (2002b). Statistical learning in infants. *Proceedings of the National Academy of Science*s, *99*, 15250-15251.

Anderson. J. L., Morgan, J. L., & White, K. S. (2003). A statistical Basis for speech sound discrimination. *Language and Speech, 46,* 155-182

Angluin, D. (1988). *Identifying languages from stochastic examples* (Tech. Rep. YALEU/DCS/RR-614). New Haven, CT: Yale University, Department of Computer Science.

Atchley, R. A., Rice, M. L., Betz, S. K., Kwasny, K. M., Sereno, J. A., & Jongman, A. (2006). A comparison of semantic and syntactic event related potentials generated by children and adults. *Brain and Language, 99,* 236-246.

Bates, E., & MacWhinney, B. (1989). Functionalism and the competition model. In B. MacWhinney & E. Bates (Eds.), The Crosslinguistic Study of Sentence Processing. Cambridge MA: Cambridge University Press.

Besson, M., & Faïta, F. (1995). An event-related potential (ERP) study of musical expectancy: Comparison of musicians with nonmusicians. *Journal of Experimental Psychology: Human Perception and Performance, 21*, 1278–1296.

Billman, D., & Knutsen, J (1996). Unsupervised concept learning and value systematicity: a complex whole aids learning the parts. *Journal of Experimental Psychology*: *Learning. Memory, and Cognition*, *22,* 458-475.

Billman, D (1989). Systems of correlations in rule and category learning: Use of structured input in learning syntactic categories. *Language and Cognitive Processes,* 4, 127-55.

Bonatti, L. L., Peña, M., Nespor, M., & Mehler, J. (2005). Linguistic constraints on statistical computations: The role of consonants and vowels in continuous speech processing. *Psychological Science, 16*, 451–459.

Brooks, L. R., & Vokey, J. R. (1991). Abstract analogies and abstracted grammars: Comments on Reber (1989) and Mathews et al. (1989). *Journal of Experimental Psychology: General, 120,* 316-323.

Chambers, K. E., Onishi, K. H., & Fisher, C. (2003). Infant lean phonotactic regularities from brief auditory experience. *Cognition, 87*, B69-B77.

Chomsky, N. (1957). *Syntactic Structures*. The Hague: Mouton & Co.

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

Chomsky, N. (1959). A review of B. F. Skinner's "Verbal behavior." *Language, 35*, 26-58.

Chomsky, N. (1995). A Minimalist Program. London: MIT press.

Christiansen, M.H. and Curtin, S.L. (1999) Transfer of learning: rule acquisition or statistical learning? *Trends Cognitive Sciences, 3*, 289–290.

Christiansen, M.H. & Chater, N. (1999). Connectionist natural language processing: The state of the art. *Cognitive Science, 23*, 417-437

Christiansen, M.H. & Chater, N. (2001). Connectionist psycholinguistics: Capturing the empirical data. *Trends in Cognitive Sciences*, *5*, 82-88.

Christiansen, M.H., Conway, C., & Onnis, L. (2007). Neural Responses to Structural Incongruencies in Language and Statistical Learning Point to Similar Underlying Mechanisms. In *Proceedings of the 29th Annual Meeting of the Cognitive Science Society.*

Connine, C., Ferreira, F., Jones, C., & Clifton, C., & Frazier, L. (1984). Verb frame preference: descriptive norms. *Journal of Psycholinguistic Research*, *13*, 307-319.

Conway, C., & Christiansen, M. H. (2005). Modality constrained statistical learning of tactile, visual, and auditory sequences. *Journal of Experimental Psychology: Learning, Memory & Cognition, 31*, 24-39.

Conway, C. M., & Christiansen, M. H. (2006). Statistical learning within and between modalities: Pitting abstract against stimulus specific representations. *Psychological Science, 17*, 905-912.

Coulson, S., King, J. W. & Kutas, M. (1998). Expect the unexpected: Event-related brain response to morphosyntactic violations. *Language and Cognitive Processes, 13*, 21-58.

Dienes Z., Broadbent D., Berry D. *(*1991*).* Implicit and explicit knowledge bases in artificial grammar learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 17*, 875-887.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14, 179-211.*

Elman, J.L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition,* 48, 71-99.

Federmeier, K. D., & Kutas, M. (1999). A rose by any other name: Long-term memory structure and sentence processing. *Journal of Memory & Language*, *41*, 469-495.

Fiser, J., & Aslin, R. N. (2002). Statistical learning of new visual feature combinations by infants. *Proceedings of the National Academy of Sciences, 99,* 15822–15826.

Frazier, L. (1987). Syntactic processing: Evidence from Dutch. *Natural Language & Linguistic Theory*, *5*, 519-560.

Frazier, L. (1990). Exploring the architecture of the language processing system. In G. T. M. Altmann (Ed.), *Cognitive Models of Speech Processing: Psycholinguistic and Computational Perspectives* (pp. 409-433). Cambridge, MA: MIT Press, Bradford Books

Frazier, L., Clifton, C., & Randall, J. (1983). Filling gaps: Decision principles and structure in sentence comprehension. *Cognition, 13,* 187-222.

Frazier, L., & Clifton, C. (1996). *Construal*. Cambridge, MA: MIT Press.

Friederici, A.D., Steinhauer, K. & Pfeifer, E. (2002). Brain signatures of artificial language processing: Evidence challenging the critical period hypothesis. *Proceedings of the National Academy of Sciences, 99,* 529-534.

Gerken, L., Wilson, R., Lewis, W. (2005). Infants can use distributional cues to form syntactic categories. *Journal of Child Language, 32,* 249-268.

Gold, E. M. (1967). Language identification in the limit. *Information and Control*, 16, 447-474.

Gómez, R. L., & Gerkin, L. (1999). Artificial grammar learning by 1-year-olds leads specific and abstract knowledge. *Cognition, 70,* 109-135.

Gómez, R .(2002). Variability and detection of invariant structure. *Psychological Science, 13,* 431-36.

Gómez, R., & Lakusta, L. (2004). A first step in form-based category abstraction by 12-month-old infants. *Developmental Science, 7,* 567-580.

Gomez, R. L., Gerken, L., & Schvaneveldt, R. (2000). The basis of transfer in artificial grammar learning. *Memory & Cognition, 28,* 253-263.

Hagoort, P., Brown, C. M., & Groothusen, J. (1993). The syntactic positive shift (SPS) as an ERP measure of syntactic processing. *Language and Cognitive Processes, 8,* 439-483.

Hare, M., McRae, K., & Elman, J. L. (2003). Sense and structure: Meaning as a determinant of verb subcategorization preferences. *Journal of Memory and Language, 48*, 281-303

Hyams N (1986). *Language acquisition and the theory of parameters*. Reidel Press.

Kelly, M. H., & Bock, J. K. (1988). Stress in time. *Journal of Experimental Psychology*: *Human Perception and Performance, 14*, 389-403.

Janata, P. (1995). ERP measures assay the degree of expectancy violation of harmonic context in music. *Journal of Cognitive Neuroscience, 7*, 153-164.

Kelly, M. H. (1992). Using sound to solve syntactic problems: The role of phonology in grammatical category assignments. *Psychological Review, 99,* 349-64.

Kelly, M. H., & Martin, S. (1994). Domain-general abilities applied to domain-specific tasks: Sensitivity to probabilities in perception, cognition, and language. *Lingua, 92*, 105-140.

Keidel, J. L. Jenison. R. L., Kluender, K. R., & Seidenberg, M. S. (2007). Does grammar constrain statistical learning? Commentary on Bonatti, Pen˜a, Nespor, and Mehler (2005). *Psychological Science, 18*, 922-923.

Koelsch S, Gunter, T., Schroeger, E., & Friederici A. D. (2003). Processing tonal modulations: an ERP study. *Journal of Cognitive Neuroscience, 15*, 1149-1159.

Koelsch, S., Gunter, T., Wittfoth, M., & Sammler, D. (2005). Interaction in syntax processing in language and music: an ERP Study. *Journal of Cognitive Neuroscience, 17*, 1565-1577.

Koelsch, S., & Siebel, W. (2005). Towards a neural basis of music perception.*Trends in Cognitive Sciences, 9*, 78-584.

Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, 307,161-163.

MacDonald, M. C., & Christiansen, M. H. (2002). Reassessing working memory: A comment on Just & Carpenter (1992) and Waters & Caplan (1996). *Psychological Review, 109*, 35–5

MacDonald, M. C., Pearlmutter, N.J., & Seidenberg, M.S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review, 101*, 676-703.

Manzini R, & Wexler K (1987). Parameters, Binding Theory and Learnability. *Linguistic Inquiry, 18*, 413-444.

Marslen-Wilson, W. D., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, *10*, 29-63.

Marcus, G. F. (1999). Do infants learn grammar with algebra or statistics? *Science*, *284*, 436-437.

Marcus, G. F. (2001). *The Algebraic Mind: Integrating Connectionism and Cognitive Science*. Cambridge, MA: MIT Press.

Marcus, G. F., & Berent, I. (2003). Are there limits to statistical learning? *Science*, *300*, 52-53.

Marcus, G. F., Vijayan, S, Bandi Rao, S., & Vishton, P. M. (1999). Rule learning by 7 month-old infants. *Science, 283*, 77-80.

Mathews, R.C., Buss, R.R., Stanley, W.B., Blanchard-Fields, F., Cho, J.R., & Druhan, B. (1989). Role of implicit and explicit procesess in learning from examples. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*, 1083-1100.

Mattys, S, & Juszcyk, P. (2001). Do infants segment words or recurring contiguous patterns? *Journal of Experimental Psychology: Human Perception and Performance*, *27*, 644-655.

Maye, J., Weker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*, B101-B111.

Maye, J., & Weiss, D. (2003). Statistical cues facilitate infants' discrimination of difficult phonetic constrasts. In B. Beachley (ed.), *BUCLD 27 Proceeding*. MA: Cascadilla Press

McKinnon, R., & Osterhout, L. (1996). Constraints on movement phenomena in sentence processing: Evidence from event-related brain potentials. *Language and Cognitive Processes, 11*, 495-523.

McRae, K., M., Hare, M., Elman, J. l., & Ferretti, T. (2005). A basis for generating expectancies for verbs from nouns. *Memory and Cognition*, *33*, 1174-1184.

McRae, K., Spivey-Knowlton, M.J., & Tanenhaus, M.K. (1998). Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language, 38,* 283-312.

Neville, H. J, Nicol, J. L., Barss, A., Forster, K. I., & Garrett, M. F. (1991). Syntactically based sentence processing classes: Evidence from event related brain potentials. *Journal of Cognitive Neuroscience, 3,* 151-165.

Onnis, L., Monaghan, P., Richmond, K., & Cater, N. (2005). Phonology impacts segmentation in online speech processing. *Journal of Memory and Language, 53,* 225-237.

Onnis, L., Monaghan, P., Christiansen, M. H., & Chater, N. (2004). Variability is the spice of learning, and a crucial ingredient for detecting and generalizing in nonadjacent dependencies. *Proceedings of the 26th Conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum Associates

Osrerhout, L., & Mobley, L. A. (1995). Event-related brain potentials and language comprehension. M. D. Rugg and M. G. H. Coles (Eds), *Electrophysiology of Mind* (pp. 171-215). Oxford University Press.

Osterhout, L., & Holcomb, P. (1992). Event-related brain potentials elicited by syntactic anomaly. *Journal of Memory and Language, 31*, 785-806.

Osterhout, L., Holcomb, P., & Swinney, D. A. (1994). Brain potentials elicited by garden- path sentences. *Journal of Experiment Psychology: Learning, Memory, & Cognition*, 20, 786-803.

Osterhout, L., McKinnon, R., Bersick, M., & Corey, V. (1996). On the language-specificity of the brain response to syntactic anomalies: Is the syntactic positive shift a member of the P300 family? *Journal of Cognitive Neuroscience, 8*, 507-526.

Osterhout, L., & Nicol, J. (1999). On the distinctiveness, independence, and time course of the brain responses to syntactic and semantic anomalies. *Language and Cognitive Processes, 14*, 283-317.

Palmer, F. R. (1981). *Semantic*s. Cambridge: Cambridge University Press.

Patel, A. D., Gibson. E., Ratner. J., Besson, M., & Holcomb, P. J. (1998). Processing syntactic relations in language and music: An event-related potential study. *Journal of Cognitive Neuroscience, 10*, 717-733.

Patel, A. D. (1998). Syntactic processing in language and music: difference cognitive operations, similar neural recourses? Music Perception, 16, 27-42.

Patel, A. D. (2003) Language, music, syntax and the brain. *Nature Neuroscience*. *6*, 674-681.

Peña, M., Bonatti, L. L., Nespor, M., & Mehler, J. (2002). Signal-driven computations in speech processing. *Science, 298,* 604–607.

Perruchut, P. Tyler, M. D., Galland, N., & Peereman, R. (2004). Learning nonadjacent dependencies: No need for algebraic-like computations. *Journal of Experimental Psychology: General, 133*, 573-583.

Pinker, S. (1984). *Language Learnability and Language Development*. Cambridge, MA: Harvard University Press.

Reber, A. S. (1989). Implicit leanring and tacit knowledge. *Journal of Experimental Psychology: General, 118*, 219-235.

Roeper T, and Williams E (1987). *Parameter Setting*. Dordrecht: D. Reidel.

Rossi, S., Gugler, M. F., Hahne, A., & Friederici, A. D. (2005). When word category information encounters morphosyntax. *Neuroscience Letters, 384*, 228-233.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature, 323*, 533 - 536

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, 35, 606-621.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science, 274*, 1926-1928.

Saffran, J. R., Johnson, E.K., Aslin, R.N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition, 70*, 27-52.

Saffran, J. R (2001a). Words in a sea of sounds: The output of statistical learning. *Cognition, 81*, 149-169.

Saffran, J. R (2001b). The use of predictive dependencies in language learning. *Journal of Memory and Language, 44*, 493-515.

Saffran, J. R (2002). Constraints on statistical language learning. *Journal of Memory and Language, 47*, 172-196.

Seidenberg, M. S., MacDonald, M. C., & Saffran, J. R. (2002). Does grammar start where statistics stop? *Science, 298*, 553-554.

Seidenberg, M.S., MacDonald, M.C., & Saffran, J.R. (2003). Response to Marcus and Berent. *Science, 300*, 53.

Skinner B F (1957). *Verbal behavior*. New York: Appleton-Century-Crofts.

Steinbeis, N., & Koelsch, S. (2007). Shared neural resources between music and language indicate semantic processing of musical tension-resolution patterns. *Cerebral Cortex, 18*, 1169-1178.

Tanenhaus, M.K. & Trueswell, J.C. (1995). Sentence Comprehension. In Eimas & Miller (Eds.) *Handbook in Perception and Cognition*, Volume 11: Speech Language and Communication (pp. 217-262). Academic Press.

Taraban, R., & McClelland, J. L. (1998). Constituent attachment and thematic role assignment in sentence processing: Influence of content-based expectations. *Journal of Memory and Language, 27*, 597-632.

Tyler, L.K., & Marslen Wilson, W.D. (1977). The on-line effects of semantic context on syntactic processing. *Journal of Verbal Learning and Verbal Behavior, 16,* 683-692.

Trueswell, J.C., Tanenhaus, M.K., & Kello, C. (1993). Verb-specific constraints in sentence processing: Separating effects of lexical preference from garden-paths. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 528-553.

Trueswell, J.C., Tanenhaus, M.K., & Garnsey, S.M. (1994). Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Journal of Memory and Language, 33,* 285-318.

Trueswell, J.C. (1996). The role of lexical frequency in syntactic ambiguity resolution. *Journal of Memory and Language, 35,* 566-585.

van Berkum, J. J. A., Brown, C. M., Zwitserlood, P., Kooijman, V.,& Hagoort, P. (2005). Anticipating upcoming words in discourse: Evidence from ERPs and reading times. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 31,* 443-467.

Vokey, J. R., & Higham, P . A (2005). Abstract analogies and positive transfer in artifical grammar learning. *Canadian Journal of Experimental Psychology, 59,* 54-61

Wexler K, and Culicover P (1980). *Formal Principles of Language Acquisition*. MIT Press.

Wexler, K. (1982). A principle theory for language acquisition. In E. Wanner and L. R. Gleitman (eds.) *Language Acquisition: The State of the Art*. Cambridge, MA: MIT Press.

Wicha, N. Y. Y., Moreno, E. M., & Kutas, M. (2003). Expecting gender: An event related brain potential study on the role of grammatical gender in comprehending a line drawing within a written sentence in Spanish. *Cortex, 39,* 483-508.

Whittlesea, B. W., and Dorken, M. D. (1993). Incidentally, things in general are particularly determined: An episodic-processing account of implicit learning. *Journal of Experimental Psychology: General, 122,* 227-248.