Frequency of what:
How simple is the story of syntax acquisition?

Campuswide Honors Senior Thesis

Amanda Nili, June 2014

Faculty Advisor: Dr. Lisa Pearl, Cognitive Sciences

University of California, Irvine

Acknowledgements: I would like to thank: my lab-mates in the Computation of Language Laboratory for their contributions to this project in its early stages. I would also like to thank my research advisor, Lisa Pearl, for being so supportive and giving me such expert advice every step of the way.

Frequency of what: How simple is the story of syntax acquisition?

Introduction

One explanation for why certain structures in a language sound better (i.e., are more "grammatical") than others to native speakers is that the better-sounding structures have been encountered more frequently. This simple relationship between frequency and acceptability—namely, that the former is the only thing that really determines the latter—has been considered viable by the linguistics community for quite some time, although it has only recently been empirically investigated. More importantly, from the standpoint of language acquisition, since the acceptability of linguistic structures varies somewhat from language to language, these grammaticality preferences must be learned. Notably, the linguistic input that children encounter, sometimes called "motherese" or "child-directed speech," is linguistically different from the linguistic input that adults encounter (typically called "adult-directed speech"). Thus, to assess this simple explanation of how structure acceptability preferences are learned, this study builds on previous work that examined adult-directed speech and examines the relationship between structure acceptability and the frequency of structures in child-directed speech.

The rest of this paper is laid out as follows: First, we discuss the important differences between grammaticality and acceptability (and why we use the latter as a means of understanding the former). Then, we review previous research on the apparent frequency-grammaticality gap, and how researchers interpret this gap to either affirm or challenge—and propose a more sophisticated alternative to—the simple story. Next, we discuss prior work by Pearl and Sprouse (in prep.) on adult-directed input that challenges the simple story; our study builds on this pilot study by looking at child-directed speech for the reasons noted above. We then describe our methods of assessing frequency, and the important implications of these methods for a theory of syntax acquisition. Finally, we discuss our findings, which do not support the simple story, but may offer evidence for a different understanding of frequency that often relies on less shallow representations.

1. Grammaticality and acceptability

With the aim of assessing grammaticality of structures, we have used acceptability judgment data collected by Sprouse and Almeida (2012). These acceptability scores are intended to represent an individual's grammar of the language (i.e. North American English), although Pearl and Sprouse (2013) note that such acceptability scores include additional factors, and are not limited to the theoretical grammar of that language.[1] Rather, these scores are a reflection of linguistic behavior—naïve native speakers are revealing preferences about actual linguistic knowledge, which should have a basis in the grammar of the language studied as the grammar is meant to be a description of native speaker linguistic knowledge. Notably, there can be differences between what is grammatical and what is acceptable. For example, consider the phenomenon of center

[1] Pearl and Sprouse (2013) note that these "other factors… include semantic plausibility, lexical properties, and parsing difficulty." The measures designed to prevent semantic interference with the data's reflection of grammaticality are discussed in section 3.

embedding of relative clauses in North American English. The grammar of this language permits center embedding of relative clauses (e.g. "The penguin *that Elise saw* was cute."), and theoretically permits this multiple times in a single sentence (e.g. "The penguin *that Elise who Olaf who Rachel loves knows saw* was cute."). The difference between these two examples is one of acceptability—while both are technically grammatical and demonstrate center embedding of relative clauses, the latter example (which has three center-embedded clauses), is less acceptable and typically impossible for a native speaker to follow. That is, we have trouble parsing the multiple subjects in "The penguin *that Elise who Olaf who Rachel loves knows saw* was cute.", although this structure does not violate any of the technical grammar rules of our language. We use acceptability scores as a stand-in for grammaticality, recognizing that additional factors matter for acceptability. Notably, since additional factors beyond structure do matter for acceptability, we might expect that if a simple relationship exists between structural frequency and grammaticality, we should see this relationship reflected in acceptability scores as well.

2. The simple and not-so-simple stories

The simple story asserts that our brains are doing very little other than simple statistical analyses with the base frequencies of the utterances we hear. This theory is not an empirically supported one; in fact, the simple story of syntax acquisition is more of a hypothesis, one with very important implications for the methods by which linguistic research is conducted.

The not-so-simple story can be instantiated by several different theories about why the base frequencies of utterance occurrences in a corpus do not match how acceptable speakers find these utterances. These theories look at more complicated ways of studying frequency, and explain the appearance of what Kempen and Harbusch (2005) call a "frequency-grammaticality gap" by offering different—typically more abstract—accounts for the ways that our brains process the structures we hear.

Below, we expand on the core aspects of the simple and not-so-simple stories for the relationship between frequency and acceptability of utterances, and review samples of literature indicating which one different researchers believe.

2.1. The simple story

The simple story is that our brains take minimally abstracted information from all linguistic input, and calculate a frequency score from that information to determine how "good" we think the utterances are that we hear. The simple theory is that base frequency of structures correlates perfectly with acceptability of those same structures. If this theory were true, then we could create plots like Figures 1 and 2, where frequency is plotted against acceptability, and expect a perfect positive linear correlation, as shown in Figure 1, so that data would only appear in Quadrants I and III, as in Figure 2. In particular, low frequency utterances (-Frequency) should have low acceptability (-Acceptability) while high frequency utterances (+Frequency) should have high acceptability (+Acceptability).

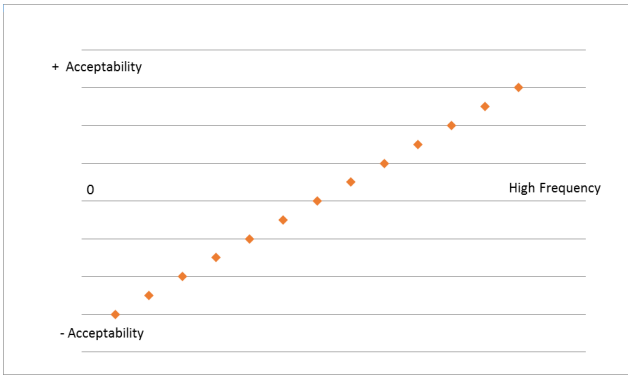

Figure 1

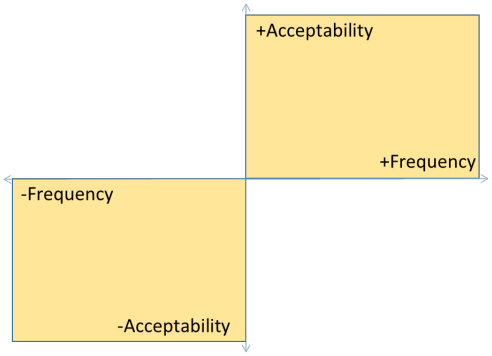

Figure 2

There is more to the simple theory than this theorized relationship (where $r = 1$). The most important component of this theory is not its prediction of what the relationship looks like, but how it defines the two things it makes predictions about. There are multiple ways of considering the frequency of structures in corpora, and there are many ways to search for frequency in these datasets. The simple theory looks at surface forms of structures (i.e. those that we can collect without having to abstract too much information from the utterance). The most superficial approach might be that that our brains collect individual words, and tabulate the frequency with which one word follows another in exactly the order the words appear in a given utterance. Were this version true, we would find utterances acceptable only because we had heard that exact series of words before. To illustrate: "The dog barks" would be acceptable only because we have heard "The dog barks" before, and many more times than "The dog speaks." To be sure, this is one example in which the semantics matter too (i.e. barking is an action exclusive to dogs and seals when not speaking figuratively, while speaking is an action we do not often attribute to dogs).

Still, this simple story disregards the function of syntax, and the fact that we produce novel combinations of words often on the basis of what grammatical categories those words are (e.g., noun, verb). For this reason, and because no authors reviewed in this paper assert that frequency of individual words is responsible for utterance acceptability, we investigate a slightly more sophisticated version of the simple story that abstracts words into their grammatical categories. For example, "The dog barks" becomes a simple subject comprised of a definite article and singular noun, followed by an intransitive verb (i.e., $\text{det}_{\text{def}} + \text{noun}_{\text{sg}} + \text{vb}_{\text{intrans}}$). Such a surface form is a way of studying syntax that incorporates fairly little abstraction.

2.2. Who believes it and who doesn't

The simple story of syntax acquisition—that we analyze the most surface forms of utterances we hear (making minimal abstraction about them) and then translate the calculated frequency of those surface forms into an acceptability score—is not supported by empirical studies on the subject. What is important to note here is that all studies comparing frequency (at various levels of abstraction) and acceptability find some form of discrepancy between those data. Those who reject a simple story of syntax acquisiton account for the discrepancy by providing theories of

different language learning and production models that account for the divergent acceptability scores. In contrast, those that accept the simple story do so by criticizing the methods employed by those researchers that find a discrepancy. This is how Kempen and Harbusch (2005) assert an implicit version of the simple theory: they object to findings of a discepancy between accepability and frequency in a study by Keller (2000), in which some structures have average acceptability scores yet do not appear in the corpus. Specifically, Kempen and Harbusch accept that the structures are not in any of Keller's corpora, but they believe that the naïve native speakers who provided acceptability scores did so erroneously, scoring utterances as ungrammatical that were in reality acceptable (Kempen & Harbusch 2005). This bad-method accusation is an important problem for the field of linguisics, as naïve native speakers are fairly popular sources for acceptability data (Sprouse & Almeida 2012). Furthermore, as discussed in section 1, the difference between grammaticality and acceptability makes it more useful to study the acceptability data that speakers can provide—who better to ask about how naïve native speakers judge linguistic data than those speakers?

Below, we explore a few alternative explanations for the gap between frequency and acceptability. Unlike the Kempen and Harbusch (2005) account, these studies offer explanations of syntax acquisition which account for the empirical data, rather than arguing with its method of collection.

2.2.1. Not-so-simple example story: Keller 2000

Keller (2000)—in a study of various grammatical constraints and their violations in English, German, and Greek—finds a gap between the frequency values of produced structures and the acceptability data collected. Keller's data are not limited to a single syntactic phenomenon, but instead cover a broad range of cross-linguistic grammar contraints. These constraints are tested in the "correct" (i.e. unviolated) form, in addition to testing the acceptability and frequency of their violated counterparts. For example, a test of Subject-Verb Agreement would look something like this:

(1) Trish has painted a picture of Arthur.
(2) *Trish have painted a picture of Arthur.

In this example, Utterance 1 represents a correct version, while Utterance 2 is the violated form. Keller's participants make acceptability judgments about both of these utterances, and Keller sorts different constraints according to how unaccepabtle the violation of each constraint is.

Keller accounts for the "gap" between unviolated (high acceptable) structures and low corresponding frequency as evidence for Linear Optimality Theory, a theory that groups contraints into types and assumes a gradient grammaticality dependent on something more than

the base frequencies at which a strucutre is produced[2]. Linear Optimality Theory (Keller 2000) is the theoretical accounting for discrepancies between acceptability and frequency data by sorting grammatical constraint into two types, or varying significance when violated: "hard" and "soft" constraints. To account for the apparent lack of relation between acceptability and frequency, Keller posits a "soft constraint" that can be violated and receive a lower acceptability score, but will still be produced somewhat frequently. One example of this is Definite Article Use (the use of indefinite articles when a definite article would be more appropriate is a violation of the soft constraint).

Definite Article Use

(3) Which friend has Trish painted a picture of?
(4) *Which friend has Trish painted the picture of?

Although this is not a violation of a hard constraint according to Keller, the asterisk notation is used to clarify which example violates (with mild acceptability) the constraint discussed here. Clear notation is necessary because soft constraint violations are not as offensive as those of hard constraints (as in 2 above). However, Keller finds that the utterance in (4) is less acceptable to naïve native speakers.

In contrast, a hard constraint accounts for low frequency utterances that are unacceptable: these are the structures that are rarely, if ever, produced. Subject-Verb Agreement (featured on the previous page) according to Keller, is a hard constraint. The constraint violation of (2) in this example should be much more obvious to the reader, as it is to participants in Keller's study. This constraint was found to be "hard"—i.e. resulting in great unacceptability when violated—cross-linguistically. In fact, Keller's study finds minimal cross-linguistic variation of hard constraints.

2.2.2. The not-so-simple story: Response to methodological critiques

As alluded to at the beginning of section 2, Kempen and Harbusch's (2005) claim of bad methodology creating a gap between frequency and acceptability has important implications for the field of linguistics research. Claims against methodology are not new, nor are they yet resolved; debate over rigor in experimentation is an ongoing topic of interest in this field (Hickok 2010). Specifically, the debate over how reliable methods of collecting acceptability (i.e. "judgment") data is over half a century old (Sprouse & Almeida 2012). In defense of linguistic methodology, Sprouse and Almeida (2012) investigate both formal and informal methods of collecting acceptability data, and the two methods share results at a 98% overlap rate.

[2] The production frequency of a structure is relevant to this study as a means of examining the linguistic input a child receives. For our purposes, "production" implies frequency of linguistic perception of those same structures, and so frequency of production will be used to discuss linguistic input in the later sections of this paper. However, the age of audience for which speech is produced (i.e. adult-directed versus child-directed) is further explored in section 4 of this paper.

This finding that the method is in fact good contradicts the Kempen and Harbusch (2005) account for a gap (i.e. the method is bad). That explanation for discrepancies between frequency and acceptability scores, based on an assumed simple relationship between these two variables, does not hold up to empirical support for the methods in question.

2.2.3. Another not-so-simple story: Jurafsky (2002) & Featherston (2004)

An alternative to questioning the methods is accounting for the gap by re-examining the theory. This is the approach that Jurafsky (2002) takes when accounting for what he calls a "mismatch" between the variables. Jurafsky's account is a probabilistic model of production, in which frequency is treated as something much more complicated than the base frequencies of parts of speech appearing in a certain order (i.e. the simple account of syntax learning). By this account, the gap found in empirical study of frequency of a structure and its acceptability score is the result of acceptability data being abstract and frequency scores being concrete in comparison. To provide an explanation for this gap, linguists would need to find a way to translate between the acceptability and the frequency data. This would then allow the two very different kinds of data to be fairly compared (Jurafsky 2002).

A probabilistic model of syntax that offers an additional explanation of the low correlation between frequency and acceptability is Featherston's (2004) model of structure selection. By this account, each potential syntactic form competes with other potential forms (each an alternative way of expressing the same thought, but all of varying grammaticality), for selection by the speaker. According to this model, the best structure is not always chosen, so frequency of a structure's appearance in a corpus cannot be expected to correlate perfectly with the acceptability score of that same structure—the two measures of the structure are not related, because some additional factor influences whether a given structure is produced (i.e. acceptability does not perfectly determine what structures are produced by speakers, or, the frequency with which structures are produced does not exclusively determine how acceptable that structure is).

2.3. Investigating the simple story for *wh*-dependencies

Pearl and Sprouse (2013) compare frequency and acceptability data using English to study a specific syntactic phenomenon: *wh*-dependencies. Studying these kinds of structures (and violations of the rules by which they abide), requires fairly abstract representation of structure. For example, the *wh*-dependency "Who did she like?" is abstracted to representations involving phrases such as complementizer phrases (CP), inflectional phrases (IP), and verb phrases (VP), as in (5). For Pearl & Sprouse (2013), the *wh*-dependency is described by the sequence of these phrases between the *wh*-word and the gap (the place where the *wh*-word is interpreted).

(5) [$_{CP}$ Who did [$_{IP}$ she [$_{VP}$ like _]]]?

wh-dependency structure = IP-VP

Thus, to study *wh*-dependencies, a search operates at the phrasal level of syntax. This higher level of abstraction is important to consider, because the story might still be "simple" in the sense that frequency of a structure determines how acceptable it is, if we define structure as something more abstract than grammatical categories. This would be the conceptualization of structure according to generative grammar. Nonetheless, it still does not offer the kind of perfect correlation that a simple theoretical account predicts: Pearl & Sprouse (2013) also find a gap between adult acceptability data and frequencies collected from adult-directed speech for the *wh*-dependencies they investigate.

3. The frequency-acceptability gap in adult-directed speech

In previous investigations of native speaker intuitions about language structure, researchers have used both formal (magnitude estimation) and informal (yes-no) acceptability judgment data because of how easily this kind of information can be collected and utilized in an assessment of grammar (Sprouse & Almeida 2012). While the methods of collecting acceptability judgment data have been attacked for a lack of proven reliability, a methodological defense of these collected judgments is possible. Sprouse and Almeida (2012) find that both methods are very nearly equal in terms of reliability[3]; however, not all methods are equally appropriate for every study. All acceptability scores used in the present study were collected by the magnitude estimation method, in which participants are asked to assign relative ratings of acceptability to each utterance. This is a preferable method for the current study because the theory tested here is one of gradient grammaticality, where some grammatical structures are more or less grammatical than other grammatical structures. Although the yes-no method can give reliable data, and acceptability judgments do not differ significantly from those scores acquired by magnitude estimation, the logic of yes-no methodology is binary and ignores the gradience that we intend to study (Sprouse & Almeida 2012; Featherston 2004). To assess gradient grammaticality, we must use data that were collected with the possibility of relative scoring. All acceptability scores used in this study are z-scores: they indicate the better-or-worse-than quality of every structure assessed, and are either above or below a mean value of acceptability. To control for semantic influence and assess the structure's grammaticality specifically, structures are assessed multiple times, in multiple instantiations; each z-score is calculated from the average score of those multiple assessments. This allows us to eliminate, to some extent, semantic interference (although the influence of semantics becomes relevant in our later discussion of the *kind* of frequency that we study).

3.1. Prior work: A pilot study by Pearl & Sprouse (in prep.) on adult-directed input

[3] Discrepancies between yes-no and magnitude estimation methods occur at a rate of approximately 2%; for Sprouse and Almeida (2012), a discrepancy occurred any time that yes-no methods reported a different score from that of the magnitude estimation (the assumption of the study being that magnitude estimation methods are more valid).

The present study follows up on previous work by Pearl and Sprouse (in prep.) that investigated adult-directed speech and studied the relationship between frequency and acceptability for a larger range of structures, rather than a particular subset of structures—e.g. *wh*-questions (Pearl & Sprouse 2013). Pearl & Sprouse (in prep) uses the same adult judgment data that we will use here, which comes from Sprouse and Almeida's (2012) work with Adger's *Core Syntax*, a linguistics textbook. This range of structures allows a broader assessment of the veracity of the simple story: if syntactic inutitions about acceptability are acquired by how frequently we perceive minimally abstracted versions of linguistic structure, then this should occur for all syntactic phenomena.[4] While Pearl and Sprouse (in prep.) have not completed a formal assessment of their data, preliminary analysis suggests that the simple story of syntax acquisition does not hold up when structure frequencies in adult-directed speech are compared with the adult judgment data (i.e. acceptability scores) of those same structures. Instead, the authors find that the frequency-acceptability gap occurs in adult-directed speech, and when studying a large range of syntactic structures.

4. A need for frequency data from child-directed speech corpora

Previous work by Pearl and Sprouse (in prep.) demonstrates the lack of empirical support for a simple account of syntax acquisition when studying adult-directed speech. However, it is important to analyze the actual input for children during the period of language acquisition if we want to understand how we come to have our grammaticality intuitions. This is particularly important given that there are several known differences between adult-directed and child-directed speech.

We examine these differences—and their relation to the simple story of syntax acquisition—in the following sections. We also describe the corpora of child-directed speech that we use to study frequency in child-directed speech.

4.1. Differences between child-directed and adult-directed speech

One common form of child-directed speech, "motherese," has many apparent differences from adult-directed speech. These differences occur at phonetic (e.g. emphasized vowel sounds) (Fernald 1985), lexical (e.g. increased use of monosyllabic words) (Yang 2004), and shallow syntactic levels (e.g. use of Determiner+Noun constructions, rather than pronouns) (Furrow 1978). Given these known differences, one potential explanation for the frequency-grammaticality gap that would still support the simple story is that child-directed speech also consists of different frequencies of structure occurrences which correlate better with acceptability data (Kempen & Harbusch 2005). In order to gain a better understanding of how useful the simple story is as a theory of syntax acquisition, it is imperative to determine if the

[4] Notably, the discussion of previous work in section 2 demonstrates that the simple story is not supported in studies of other languages, or smaller subsets of structure in North American English. However, it may be that these were exceptional cases, rather than the rule.

frequency-acceptability gap, a notable problem for this simple story, exists when frequencies are collected from the actual input that children receive, rather than the input adults receive and produce. Pearl and Sprouse (2013) compare child-directed speech and adult-directed speech at a more abstract syntactic level of representation for *wh*-questions (involving phrasal structure) and interestingly find little difference at this level of abstraction.

Additional support for the marked difference between child-directed speech and adult-directed speech comes from the Vygotsky's theory of the *zone of proximal development* (Vygotsky 1978). This developmental psychology account is one of scaffolding; child-directed speech responds to the child's level of acquisition by adjusting the level of difficulty to accommodate what the child has learned and still needs to learn. Thus, it is considerably different from adult-directed speech. The application of Vygotsky's theory to syntax acquisition could mean that frequencies of most structures studied should be very different from those that Pearl and Sprouse (in prep.) collect in adult-directed speech, at least when comparing the data for more complex syntactic structures.

Interestingly, Yang (2004, 2011) provides potential evidence that there does *not* seem to be a frequency-acceptability gap when the learner tracks the frequency of very abstract structural representations (called linguistic parameters). The frequency of unambiguous "linguistic signatures" for these structural representations correlates very well with how early children learn the structural representations. If we assume that the earlier a structure is learned, the earlier a child finds that structure acceptable, then this indicates that frequency correlates quite well with acceptability. The key is the frequency of *what*, a question we return to in section 7.1.

Notably, our approach investigates structural representations that are less abstract than Yang's linguistic parameters. In particular, the current study examines the large range of structures presented in Adger's *Core Syntax* (Sprouse & Almeida 2012), and thus seeks to account for a relationship (or lack thereof) between the frequencies of different types of syntactic phenomena at a very shallow level, and their associated acceptability scores.

4.2. Corpora from CHILDES

The child-directed speech corpora consist of the following portions of the CHILDES datasets that have speech directed at children of the ages indicated:

Corpus Name	Age Range	Number of Utterances
Brown-Adam	2;3-4;10	26,280
Brown-Eve	1;6-2;3	14,245
Brown-Sarah	2;3-5;1	46,948
Soderstrom	0;6-1; 0	21,334
Suppes	1;11-3;11	35,906
Valian	1;9.20-2;8.24	25,550

All ages are within an appropriate period for studying language acquisition. While this study makes no comment on the debate between a critical and a sensitive period of language

acquisition, the age range does not approach (or surpass) any of the barrier ages implicated in either theory (Newport 2002). A total of 170,263 utterances from these corpora are mined for frequency data in order to provide a large enough sample size for our work to have meaningful implications in the field of language acquisition research.

5. Methods of assessing frequency

To assess the frequency of each structure in the corpora, it is necessary first to determine how to annotate and search for each structure. For the purposes of an investigation into the simple story of syntax acquisition, it is important that our queries involve minimal abstraction made at the structural level; meanwhile, each of these searches must capture what is most important about the structure (e.g. a subject-verb agreement violation or an intransitive verb). To identify the relevant structures in child-directed speech, the structures can be translated into queries written for Tregex (a linguistic pattern-matching utility: http://nlp.stanford.edu/software/tregex.shtml). These queries are intended to search for those salient lexical and syntactic features of the sentences previously scored for acceptability.

In the sections below, we discuss how queries were formulated (and the implications of our annotations for the kind of frequency assessed), as well as how the raw frequency values are normalized and smoothed to allow for accurate visual comparison between frequency and acceptability scores.

5.1. Queries

All queries are created to search a dataset of utterances annotated independently of acceptability studies (i.e. the annotation used for the CHILDES database does not mark a violation of Subject-Verb Agreement as such; instead, for example, we must search for all plural nouns followed by a singular verb form).

As an illustration of the ways in which we annotate utterances, making minimal abstraction about their structures, we provide the following examples:

(6) A sentence that received a high acceptability score: “The scissors are lost.”

[S [NP [DT The] [NNS scissors]] [VP [AUX are] [VBN lost]]]

The salient features of this sentence (S) are the noun phrase (NP), which is headed by a definite article “the” (DT) and contains a plural noun (NNS), and the verb phrase (VP) which consists of an auxiliary "are" (AUX) and past-participle "lost" (VBN).

The query design for this structure is: (/S/ <1 (/NP/ < /NNS/) <2 (/VP/ < (/AUX/ . /VBN/))), where we look for a sentence that has an NP with a plural noun as its first child and a VP with an auxiliary verb immediately followed by a participle. This would identify sentences such as “The scissors are lost” and “The scissors is lost”. We then hand-check the identified sentences to see if

they follow the salient aspects of the structure we want. In this case, “The scissors are lost” would while “The scissors is lost” would not.

(7) A sentence that received a low acceptability score: “The pigs grunts.”

[S [NP [DT The] [NNS pigs]] [VP [VBZ grunts]]]

The salient feature of this sentence (S) is its subject-verb agreement error. In this particular structure, the subject-verb agreement error consists of a noun phrase (NP) that includes a plural noun (NNS), followed by a verb phrase (VP) that includes the third-person singular verb form (VBZ).

The query design for this structure is: /S/ < ((/NP/ < NNS) $+ (/VP/ <` VBZ)), where we look for a sentence that has a NP with a plural noun as one of its children, and a VP with the verb form appropriate for a singular noun (in the third person). This would identity sentences such as “The pigs grunts.” and exclude sentences like “The pig grunts.” or “The pigs grunt.” To ensure that all identified forms were instantiations of subject-verb agreement errors (and not erroneously annotated), we hand-check the results here, as well.

To ensure that queries are correctly written, we use a test file containing correctly parsed versions of the sample sentences, against which every query design is checked (e.g., if a query is written correctly, it should always find at least one result in the test file).

5.1.1. Query translation

Pearl and Sprouse (in prep) created Tregex queries for identifying these structures in adult-directed speech. However, the electronic corpora containing the adult-directed speech samples are formatted differently than the electronic corpora used for the child-directed speech samples (e.g., they use different syntactic category labels and different structural notation). The annotations of adult-directed speech include labels for subject and object within a sentence, as well as the traces left by nouns within complementizer phrases, *wh*-dependencies, and passive voice. This information is more abstract than that of the child-directed annotations, which differentiate between nouns only in terms of number (i.e. singular or plural). In the child-directed annotations, “Pigs love truffles.” is a string of *Plural noun + transitive verb + plural noun*, and there is no account of “truffles” as a plural noun object. In the adult-directed utterances, this information is accounted for, as all object nouns are labeled accordingly, and queries can search for that information in the corpora. At the more abstract (i.e. structural) level of passive voice, the trace that “Most horses” leaves in “Most horses have been domesticated $t_{most\ horses}$.” is not annotated in the child-directed corpora. Instead, in the child-directed corpora, “have” is annotated in the same way as every other auxiliary verb, and the complex syntactic relationship between domestication and the horses (which appear sentence-initially but are still objects of the verb phrase) is disregarded in this search. Thus, a much less abstract search is performed.

All queries are rewritten using the appropriate annotation (i.e. that used for the child-directed speech data).

5.2. Calculating frequency values

Once frequency values were collected by the methods described in sections 3.1 and 5.1, we normalized these frequencies (e.g., a structure of raw frequency 20, divided by the total number of utterances in the corpora 170,263, becomes 0.000117465). However, these normalized frequencies were very small numbers (e.g., 0.000117465), and many of the structures were not found at all in the corpora (i.e. they had frequencies of zero). We smoothed the data by adding +0.5 to all normalized frequency scores (e.g., 0 became 0.5, and 0.000117465 became 0.00012042), so that, when taking $\log_{10}$ of the normalized frequency scores, we would not have to take the $\log_{10}$ of zero (which is undefined). It was necessary to take $\log_{10}$ of all the values in order to graph these very small numbers against their much larger acceptability score counterparts. This results in larger absolute values (although they all become negative numbers as a result), e.g., $\log_{10}$ (0.000117465) = -3.930090286.

6. Results

After collecting and adjusting the frequency scores, we entered these data into an Excel file, to allow for visual comparison of acceptability and frequency, as well as calculation of an *r* score to assess the correlation between these two kinds of data. In the sections below, we explore the different kinds of mismatches between frequency and acceptability that appear in our data.

6.1. Comparison of structure acceptability and frequency in CHILDES corpora

In Figure 3, we see an absence of the positive linear relationship predicted by the simple story (and pictorially represented in Figure 1).

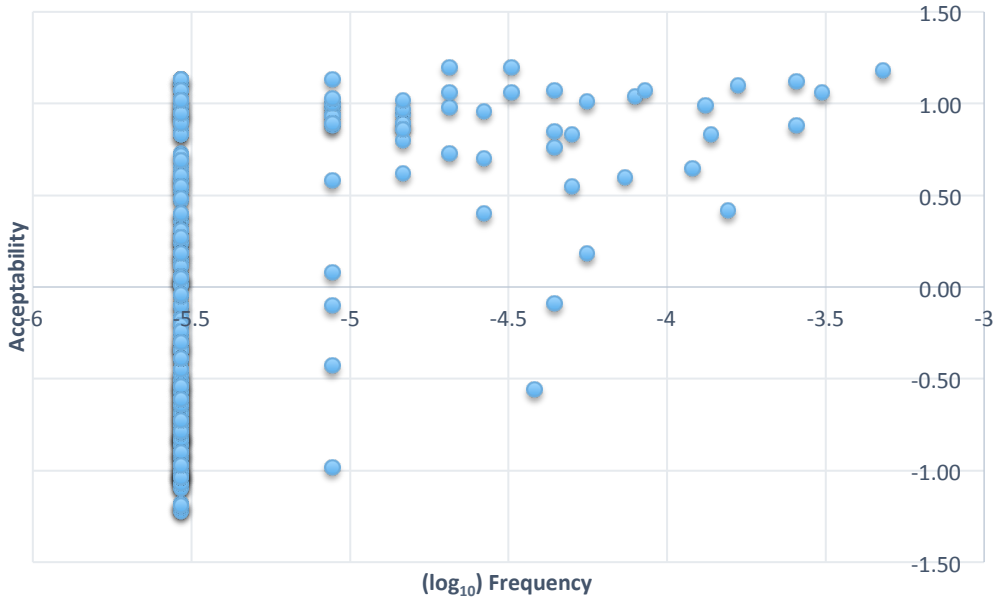

Figure 3: The x-axis represents ($\log_{10}$) frequency scores while the y-axis represents adult acceptability scores. Each point on this graph represents a single structure.

Although there are some structures for which the simple story seems to be true (i.e. their acceptability scores align with frequency values, in compliance with a positive linear relationship) there are many that do not follow this pattern. Recall Figure 2: in a graph where the x-axis is frequency (from very low to very high) and the y-axis is acceptability (also from very low to very high), the simple story predicts that data points that fall only in Quadrants I (high frequency/high acceptability) and III (low frequency/low acceptability). However, as Figure 3 demonstrates, there are many exceptions to this assumed relationship.

The following sections discuss the different types of data not predicted (or explained) by the simple story.

6.1.1. Quadrant II violations: High acceptability and low frequency

One high acceptability structure that has a low frequency value is *Subject (nominative pronoun) + "have"-auxiliary +transitive verb past participle + object (accusative pronoun)*, or "She has kissed her." The acceptability score of this structure is fairly high for the data (0.80), while its raw frequency value is very low (i.e. it only appears twice in the CHILDES corpora). This structure, like other Quadrant II violations, demonstrates that minimally abstracted structures can be infrequent but still highly acceptable.

6.1.2. Quadrant IV violations: Low acceptability and some frequency

If our knowledge of acceptability is based solely on how often we hear structures, then all unacceptable structures should have low frequencies. However, the utterance "Letter is on the table." and its structure (*NP = [singular count noun with no determiner] + verb + PP, nothing after*) have an acceptability score of -0.9 and a raw frequency of 7. Notably, this unacceptable structure has the same raw frequency value as the highly acceptable utterance "Joss's idea is brilliant." (which has an acceptability score of 1.07). Thus, this data point is unexpected under the simple story.

6.1.3. That line in Quadrants II and III: Acceptability variation in no-frequency structures

The data present with a very different kind of linear relationship than that predicted by the simple story: 168 of 219 structures searched for in the CHILDES corpora have frequencies of zero. The following examples show the considerable variation in acceptability between these structures:

(1) "Peter is pigs." is an instantiation of the structure: *Subject (name) + be + object (plural noun)*. This structure has an acceptability score of -1.20.
(2) "His analysis of her was flawed." is an instantiation of the structure: *Noun phrase (possessive + noun + PP-of-simple-noun) + verb + participle*. This structure has a low acceptability score of -0.39.

(3) "Kim should leave for work on time." is an instantiation of the structure: *Subject= (singular noun + name) + modal + intrans-verb + PP=[preposition+noun] + PP=[preposition+noun]*. This structure has a very high acceptability score of 1.13. [5]

The 168 structures that fall into this vertical line represent an *r* of 0; these structures show a lack of relationship between the frequency of minimally abstracted structures and acceptability data.

6.2. Discussion of *r* and the structures that support the simple story

The simple story predicts a perfect correlation, or an *r* of 1. The correlation between frequency values of utterances in the CHILDES corpora and corresponding acceptability data in this study is not perfect; the *r* value is 0.509997592. This is a positive relationship, but it is not a linear one. However, the *r* value is considered very strong, and some of the structures do fall into the simple story's predicted line. Figure 4 highlights three example structures that represent the predicted frequency-acceptability relationship of the simple story.

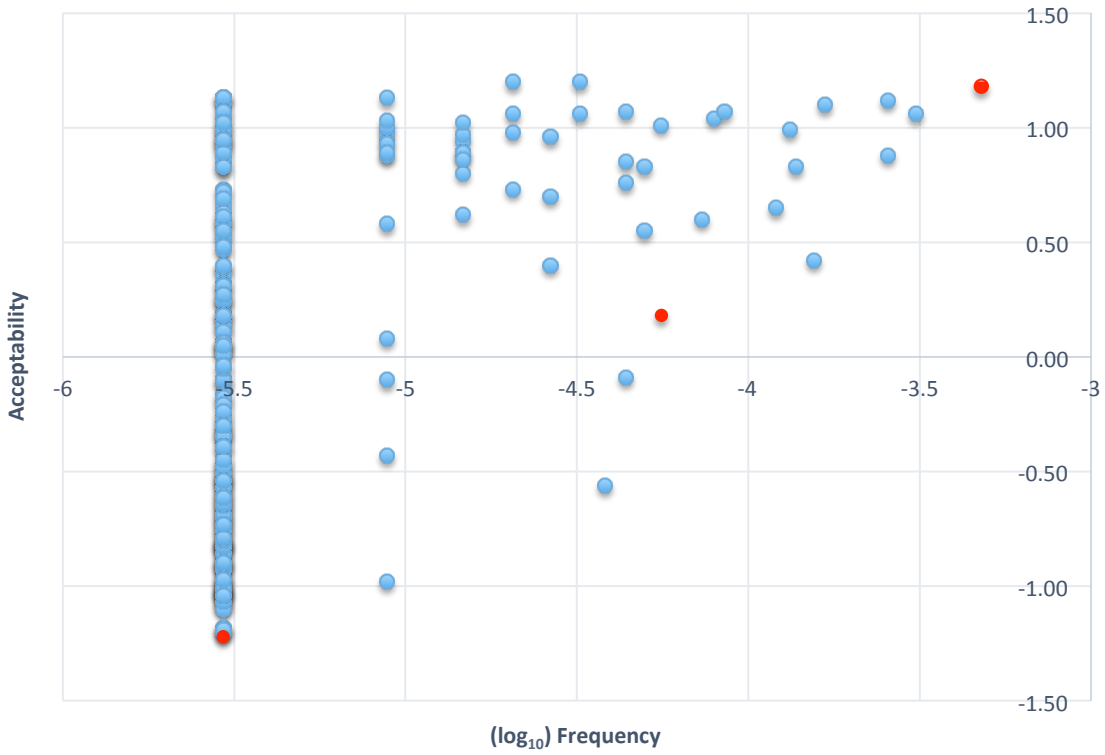

Figure 4: When the Simple Story Holds

[5] While these structures are occasionally represented here in the more abstracted versions, all queries and annotations used in this study are written with the minimally abstracted information of each structure (as stated in section 5.1 and shown in the table in Appendix A). For example, "Peter is pigs." would be searched for as a *Singular noun (name) + be-verb + plural noun*, so that object and subject status are disregarded.

In the following section, we discuss these data points that the simple story predicts.

6.2.1. Correlated scores

The unacceptable, no-frequency utterance, “It’s arrived first that Julie and Jenny”, represents an expletive-*it* located at the head of an intransitive verb phrase, with the subjects of that phrase located at the end of sentence, after complementizer-that. This is a structure not permitted by the English grammar, or (according to its acceptability score), by naïve native speakers of that language. Because it has the lowest acceptability score, this structure should also have the lowest possible frequency (zero), and it does.

The mildly acceptable (0.18), moderately frequent utterance (appearing 9 times), “I worry if the lawyer forgets his briefcase at the office” represents a simple subject, a main verb that takes an embedded-if clause with a transitive verb, a direct object, and a locational prepositional phrase. There is no obvious violation of the known rules of English grammar, and naïve native speakers score it as more acceptable than average, so this structure supports the simple story’s claims.

The most acceptable (1.18), most frequent utterance (appearing 81 times), “Genie bought the mirror” represents a subject that consists of a name, a transitive verb, and its direct object (which consists of a determiner and a noun). The fact that the most acceptable structure is also the most frequently appearing in the corpora offers some support to the simple story of syntax acquisition (as does the fact that the least acceptable structure has the lowest possible frequency). These sample structures show that the simple story is not simply wrong; however, the numerous exceptions to this story do suggest that something more complicated is happening in the process of acquiring syntax. We posit that these well-correlated structures are evidence for the simple story because surface analysis of these structures gives a learner all of the information he or she needs, without abstraction or a learning bias to offer support to raw frequency. In short, the simple story works for some structures, but apparently there are many structures that we understand in terms beyond the surface ones laid out by this simple story.

7. Discussion: Frequency of what?

7.1. Levels of abstraction

There are multiple levels of abstraction at which we can study structures in order to calculate their frequencies. Notably, certain violations can only be studied at a lower or higher level of abstraction than the one we generally adopted for the purpose of this study. For example, Semantic Category Violations such as “The book ran” do not seem unacceptable when we only account for the syntactic structure. However, on the semantic level, they sound terrible: Books, we know, do not run. We cannot adequately calculate the frequency of this structure by simply searching for a subject comprised of a determiner and a noun, followed by an intransitive verb. There is more to understanding this utterance (and its underlying structure) than its shallow, minimally abstracted structure. In our study, we account for these errors by requiring a more

abstract kind of data—any calculation for this structure only included inanimate nouns combined with animate verbs. This is an inclusion of a very abstract understanding of syntax that incorporates some semantic information, rather than the more shallow one with which we approached the other structures.

Because we do not know how abstract the analyses of children acquiring language are, it is impossible to know which level of abstraction is the appropriate one. Studies like the present one can evaluate a theory by collecting frequency data using the terms (i.e. level of abstraction) set forth by that theory. What we have done here is assess the simple story of syntax acquisition—one that claims very minimal abstraction is made about the structures children hear. With the aforementioned exception of Semantic Category Violations, we annotated and analyzed structure at a very shallow level of abstraction. As discussed in section 6.2., there are some structures for which the simple story is an adequate account. However, for those kinds of violations (i.e. the frequency-acceptability gap) discussed in section 6.1., the simple story is not saying enough. Very little of the data are accounted for by the simple story—it is more exception than rule—and for that reason, it is not an adequate account of how we acquire our intuitions about which structures are acceptable and which are not.

7.1.1. Links to the language acquisition process

Our previous discussion of Yang (2004, 2011) shows that there are alternate, and more abstract, ways of understanding syntax and the language acquisition process. Yang's (2004) findings suggest that more abstract structure frequencies correlate better with acceptability, especially when age of acquisition is considered (i.e. the sooner children learn more abstract structures, the sooner they view those structures as acceptable).

7.2. Future directions for research

We find that the simple story of syntax acquisition is an inadequate explanation of all the data; when these data are collected by sound research methods, there is not enough support for a theory of acquisition that upholds children are only analyzing the base frequencies of shallow structure. In the following sections, we discuss ways that future studies might better understand the nature of first language syntax acquisition.

7.2.1. Grammaticality according to children

Children do, at certain levels, receive different input. Prior research finds that the differences between child- and adult-directed speech are not significant at a more abstract syntactic level (Pearl & Sprouse 2013). Our preliminary investigation also shows that there are negligible differences between child- and adult-directed speech when working with much more shallow structures (Pearl & Sprouse in prep.). However, the corpora used provide a fairly wide age range (six months to five years and one month). This fact may account for the lack of evidence

supporting Vygotskian expectations of the data, which would assume different structures being used in speech directed at younger children as opposed to older children[6].

Another consideration is that, while the input is similar,[7] the interpretation of that input may be very different. Adults do not speak very differently at a structural level whether the audience is a child or another adult, but their audiences are potentially very different. As Yang (2004) notes, child learning biases can help account for an otherwise apparent poverty of stimulus; it is plausible that children have different (immature) knowledge than adults do, which affects what structural input they attend to. Child acceptability judgments could offer a more fair comparison between acceptability and (shallow) frequency scores, providing support for the simple story. More importantly, these child acceptability data could give researchers insight into how children acquire the syntax of their first languages.

7.2.2. A more sophisticated theory

The present study does not support a simple story of syntax acquisition. Shallow structure frequencies do not correlate with many adult acceptability judgments. Our literature review shows that alternative accounts for acquisition—i.e. those that study more abstract syntactic information—are able to close the frequency-acceptability gap. For this reason, we believe that a more sophisticated theory would better account for the data those data unexplained by the simple story, and that the phrase with which we have titled this larger section ("Frequency of what?") is essential to understanding the role of frequency in syntax acquisition. It is not enough to say that frequency of a structure very basically determines the acceptability of that structure; instead, future research should investigate the level of abstraction (and whether that varies between different syntactic phenomenon) necessary to account for a larger portion of the data.

References

[6] An age range has the potential to average (and thereby diminish the visibility of) any differences that occur at developmental stages, although the necessary input for a child of six months is likely different, at least at the level of syntactic structure complexity, from that provided to a child of five years.

[7] A finding which supports the validity of prior studies that used adult-directed speech only.

Crocker, M. W., & Keller, F. (2006). Probabilistic grammars as models of gradience in language processing. *Gradience in grammar: Generative perspectives*, 227-245.

Culy, C. (1998). Statistical distribution and the grammatical/ungrammatical distinction. *Grammars*, *1*(1), 1-13.

Featherston, S. (2004). The decathlon model of empirical syntax. *Linguistic evidence: Empirical, theoretical, and computational perspectives*, 187-208.

Fernald, A. (1985). Four-month-old infants prefer to listen to motherese. *Infant Behavior and Development, 8,* 181-195.

Furrow, D., Nelson, K., & Benedict, H. (1979). Mothers' speech to children and syntactic development: some simple relationships. *Journal of Child Language, 6*, 423-442. doi:10.1017/S0305000900002464.

Hickok, G. (2010, June 18). Weak quantitative standards in linguistics research? The Debate between Gibson/Fedorenko & Sprouse/Almeida [Msg 6]. Message posted to http://www.talkingbrains.org/2010/06/weak-quantitative-standards-in.html

Jurafsky, D. (2002). Probabilistic modeling in psycholinguistics: Linguistic comprehension and production. *Probabilistic linguistics*, *21*.

Keller, F. (2001). Gradience in grammar: Experimental and computational aspects of degrees of grammaticality. 1-388.

Kempen, G., & Harbusch, K. (2005). The relationship between grammaticality ratings and corpus frequencies: A case study into word order variability in the midfield of German clauses. *Linguistic evidence: Empirical, theoretical, and computational perspectives*, 329-349.

Newport, E. L. (2002). *Critical Periods in Language Development*. Retrieved from http://www.bcs.rochester.edu/people/newport/Newport-ECS-A0506.pdf

Pearl, L., & Sprouse, J. (2013). Syntactic islands and learning biases: Combining experimental syntax and computational modeling to investigate the language acquisition problem. *Language Acquisition*, *20*(1), 23-68.

Schütz, R. (2014) "Vygotsky & Language Acquisition." *English Made in Brazil.* Retrieved from: http://www.sk.com.br/sk-vygot.html

Sorace, A., & Keller, F. (2005). Gradience in Linguistic Data. *Lingua*, *115*(11), 1497-1524.

Sprouse, J., & Almeida, D. (2012). Assessing the reliability of textbook data in syntax: Adger's Core Syntax. *Journal of Linguistics*, *48*(3), 609-652.

Vygotsky, L. (1987). *Mind in Society: The Development of Higher Psychological Processes*. Cambridge, MA: Harvard University Press.

Yang, C. D. (2004). Universal Grammar, statistics or both?. *Trends in cognitive sciences*, *8*(10), 451-456.

Yang, C. D. (2011). Computational Models of Syntactic Acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science*. doi: 10.1002/wcs.1154.

Appendix A: Structures from Adger's *Core Syntax* and frequencies from CHILDES corpora.

Acceptability	(log_{10}) Frequency	Salient Structure	Instantiation of Structure from Adger's *Core Syntax* (Utterance)
0.65	-3.91936642	subj=[dt + noun-sg] + present-tense-verb-intrans-s	**2.0.1.g. The pig grunts. [plain or embedded]**
0.55	-4.301701356	subj=[dt + noun-pl] + present-tense-verb-intrans	**2.0.2.g. The pigs grunt. [plain or embedded]**
-0.81	-5.532150277	subj=[dt + noun-sg] + present-tense-verb-intrans	**2.0.3.* The pig grunt. [plain or embedded]**
-0.26	-5.532150277	subj=[dt + noun-pl] + present-tense-verb-intrans-s	**2.0.4.* The pigs grunts. [plain or embedded]**
-0.71	-5.532150277	subj=[dt + noun-pl] + be [is/was] + adjective/past-participle	**2.53.* The scissors is lost. [plain or embedded]**
1.01	-4.253396676	subj=[dt + noun-pl] + be [are/were] + adjective/past-participle	**2.53.g. The scissors are lost. [plain or embedded]**
0.36	-5.532150277	subj=pro + (optional)quant + tensed-verb + complex-obj=[subj=acc-pro + to + non-tensed-verb + adjective]	**2.68.g We (all) thought him to be unhappy.**
0.73	-5.532150277	subj=pro + (optional)quant + tensed-verb + complex-obj=[subj=nom-pro + tensed-verb + adjective]	**2.69.g We (all) thought he was unhappy.**
-0.80	-5.532150277	subj=pro + (optional)quant + tensed-verb + complex-obj=[subj=nom-pro + to + non-tensed-verb + adjective]	**2.70.* We (all) thought he to be unhappy.**
-0.83	-5.532150277	subj=pro + (optional)quant + tensed-verb + complex-obj=[subj=acc-pro + tensed-verb + adjective]	**2.71.* We (all) thought him was unhappy.**
0.96	-4.577907768	subj=[det + noun-pl] + intrans-verb-past-tense	**2.81a.g The bears snuffled.**
-1.08	-5.532150277	subj=[det + noun] + intrans-verb-past-tense+s	**2.81b.* The bear snuffleds.**
0.70	-4.577907768	PP=[prep + obj=[det + noun]] + subj=name + intrans-verb + adverb + adverb	**3.14.g At the club, Jerry danced extremely frantically.**

Acceptability	(log_{10}) Frequency	Salient Structure	Instantiation of Structure (Utterance)
0.13	-5.532150277	adverb + adverb + subj=name + intrans-verb + PP=[prep + obj=[det + noun]]	**3.15.g Extremely frantically, Jerry danced at the club.**
-1.08	-5.532150277	adverb + prep + subj=name + tensed-verb-intrans + adverb + np=[det + noun]	**3.16.* Frantically at, Jerry danced extremely the club.**
-1.04	-5.532150277	tensed-vb + adverb + subj=name + adverb + PP=[prep + obj=[det + noun]]	**3.17.* Danced extremely, Jerry frantically at the club.**
0.85	-4.356059018	subj=[det + adj + noun-sg] + intrans-verb	**3.18.g The old house collapsed.**
-1.19	-5.532150277	subj=[noun-sg + det + adj] + intrans-verb	**3.19.* House the old collapsed.**
0.94	-5.532150277	name + and + name + intrans-verb + adverb	**3.33a.g Julie and Jenny arrived first.**
-1.18	-5.532150277	expl=it + be + name + intrans-verb-past/past-participle + that + name + and + adverb	**3.33d.* It was Jenny arrived that Julie and first.**
-1.22	-5.532150277	expl-it + be + past-participle-intrans-verb + adverb + that + name + and + name	**3.34.* It's arrived first that Julie and Jenny.**
0.98	-5.055029022	subj=bare-plural + trans-verb + obj=bare-plural	**3.50.g Pigs love truffles.**
0.88	-3.592631024	subj=bare-plural + verb + to + trans-verb + obj=bare-plural	**3.51.g Humans love to eat pigs.**
-1.20	-5.532150277	subj=name + be + obj=bare-plural	**3.52.* Peter is pigs.**
0.42	-3.807874407	subj=bare-plural + verb + to + trans-verb + obj=[dem-det + plural]	**3.57.g Humans love to eat those pigs.**
0.40	-4.577907768	subj=bare-plural + verb + to + trans-verb + obj=[det + adj + plural]	**3.58.g Humans love to eat the old pigs.**
0.32	-5.532150277	subj=bare-plural + verb + to + trans-verb + obj=[det + adj + plural + clause=[relpro + modal + intrans-verb]]	**3.59.g Humans love to eat some happy pigs that can fly.**
-1.10	-5.532150277	subj=name + be + obj=[dem-det + plural noun]	**3.63.* Peter is those pigs.**
Acceptability	(log_{10}) Frequency	Salient Structure	Instantiation of Structure (Utterance)

0.08	-5.055029022	subj=[bare-plural + PP=[prep + bare-plural]] + trans-verb + to + trans-verb + obj=bare-plural.	**3.73.g Owners of pigs love to eat truffles.**
-0.77	-5.532150277	subj=[bare-plural + PP=[prep + NP=[det+noun]]] + 3rd-person-sg-verb + to + trans-verb + obj=bare-plural	**3.74.* Owners of a pig loves to eat truffles.**
0.94	-4.833180273	subj=expl-it + expl-verb	**3.77.g It rained.**
-0.85	-5.532150277	subj=[dt + noun] + expl-verb.	**3.79.* The weather rained.**
-0.88	-5.532150277	subj=name + trans-verb(obligatory)	**3.92.* Andy demonized.**
0.98	-4.687052237	subj=name + trans-verb(obligatory) + obj=name	**3.92.g Andy demonized David.**
-1.03	-5.532150277	subj=name + trans-verb(obligatory) + adjective	**3.112.* Andy demonized old.**
-0.91	-5.532150277	subj=name + trans-verb(obligatory) + PP=[prep + obj=[dt+noun]]	**3.113.* Andy demonized up the river.**
0.94	-5.055029022	subj=name + trans-verb(+speak) + obj=[dt+noun](-concrete-object)	**3.115.g Genie chanted the prayer.**
0.72	-5.532150277	subj=name + trans-verb(+speak-manner) + obj=[comp-that + subj=pro + be + adjective]	**3.116.g Genie chanted that she was tired.**
-0.98	-5.055029022	subj=name + trans-verb(+speak) + obj=[dt+noun](+concrete object)	**3.117.* Genie chanted the mirror.**
-0.56	-4.418206925	subj=dt+noun(-anim) + intrans-verb(+event)	**3.118.* The bookcase ran.**
0.83	-3.860052419	subj=dt+noun + intrans-verb	**3.118.g The thief ran.**
1.18	-3.319962673	subj=name + trans-verb + obj=[dt+noun]	**3.124.g Genie bought the mirror.**
-0.16	-5.532150277	subj=name + intrans-verb + adj (w/o PP but that must take "of")	**3.148.* Julie became fond.**
Acceptability	($\log_{10}$) Frequency	Salient Structure	Instantiation of Structure (Utterance)
-0.31	-5.532150277	subj=[wh-obj +	**3.152.g What Julie became was**

		subj=name + intrans-verb] + was + adjective + PP=[of + NP]	**fond of the book.**
-0.61	-5.532150277	subj=[wh-obj + subj=name + aux-do + PP=[of+NP]] + was + intrans-verb + adjective	**3.153.* What Julie did of the book was become fond.**
-0.49	-5.532150277	subj=[trans-verb + noun=det+n + adverb] + is + obj=[det + adj + noun + to + trans-verb]	**4.22.d.* Burn the letters quickly is the best thing to do.**
0.86	-5.532150277	subj=[trans-verb+ing + noun=det+n + adverb] + is + obj=[det + adj + noun + to + trans-verb]	**4.22.e.g Burning the letters quickly is the best thing to do.**
0.93	-5.055029022	subj=[1st-per] + trans-verb + obj=[1st-per-refl], nothing after	**4.37.g I shaved myself.**
-1.07	-5.532150277	subj=[1st-per-refl] + trans-verb + obj=[1st-pers]	**4.38.* Myself shaved me.**
0.52	-5.532150277	subj=[det+n + 1st-per-subj + trans-verb] + intrans-verb.	**4.44.g The man I saw left.**
-0.83	-5.532150277	subj=[det+n + 1st-per-subj + trans-verb] + trans-verb + 1st-per-refl-pro.	**4.45.* The man I saw shaved myself.**
-0.93	-5.532150277	subj=name + verb + PP=[P + name] + obj-pro	**4.68a.* Benjamin gave to Lee it.**
0.15	-5.532150277	subj=name + verb + complex-obj=[subj=pro + modal/aux + non-finite-verb] + conjunction + non-finite-verb + subj=pro + aux-do.	**4.69b.g Benjamin said he would run away and run away he did.**
0.89	-4.833180273	subj=name + verb + obj=pro + PP = [prep + name], nothing after	**4.69b2.g Benjamin gave it to Lee.**
0.02	-5.532150277	subj=name + verb + complex obj=[] + conjunction + non-fin-verb + simple obj + + PP=[prep+name]+ subj=[pro] + aux-do	**4.71.g Ben said he would give the cloak to Lee and give the cloak to Lee he did.**
Acceptability	($\log_{10}$) Frequency	Salient Structure	Instantiation of Structure (Utterance)
-0.45	-5.532150277	subj=name + trans-verb + modal + obj=name,	**5.08.* George seek may Isabelle.**

		nothing after	
0.88	-5.055029022	subj=name + modal + trans-verb + obj=name, nothing after	**5.08.g George may seek Isabelle.**
-0.97	-5.532150277	complex-subj=[wh-obj + subj=name + trans-verb] + is + modal + trans-verb + obj=name	**5.09.* What George does is may seek Isabelle.**
0.24	-5.532150277	complex-subj=[wh-obj + subj=name + modal + trans-verb] + is + trans-verb + obj=name	**5.09.g What George may do is seek Isabelle.**
-0.76	-5.532150277	subj=name + modal + modal + intrans-verb + pp=[prep+noun] + pp=[prep+noun]	**5.13.* Joe must should leave for work on time.**
1.13	-5.055029022	subj=name + modal + intrans-verb + pp=[prep+noun] + pp=[prep+noun]	**5.13.g Joe should leave for work on time.**
0.89	-4.833180273	subj=pro + trans-verb-past-tense + complex obj=[subj-pro + trans-verb-past-tense + predicate adjective], nothing after	**5.19.g I believed she was pregnant.**
0.06	-5.532150277	subj=pro + trans-verb-past-tense + complex obj=[subj-pro + trans-verb-present-tense + predicate adjective], nothing after	**5.21.* I believed she is pregnant.**
0.60	-5.532150277	subj=pro + trans-verb-past-tense + complex obj=[subj-pro + might-modal + trans verb + predicate adjective], nothing after	**5.25.g I believed she might be pregnant.**
-0.11	-5.532150277	subj=pro + trans-verb-past-tense + complex obj=[subj-pro + may-modal + trans verb + predicate adjective]	**5.27.* I believed she may be pregnant.**
-0.82	-5.532150277	subj=name + modal + trans-verb-tensed + obj=name	**5.31.* Dale might loved Clare.**
Acceptability	($\log_{10}$) Frequency	Salient Structure	Instantiation of Structure (Utterance)
-1.00	-5.532150277	subj=name + do-aux + tensed-trans-verb + obj=name	**5.36.* Dale do loved Clare.**

0.03	-5.532150277	sent1=name + trans-verb + complex obj=[pro-subj + modal + intrans verb] + sent2 = [intrans-verb + pro-sub + modal]	**5.37.g Benjamin said he would run away and run away he will.**
-0.83	-5.532150277	sent1=name + trans-verb + complex obj=[pro-subj +in trans verb] + sent2 = [intrans-verb + pro-subj]	**5.38.* Benjamin said he ran away and ran away he.**
-0.19	-5.532150277	sent1=name + trans-verb + complex obj=[pro-subj + intrans verb] + sent2 = [intrans-verb + pro-subj + do-aux]	**5.39.g Benjamin said he ran away and ran away he did.**
1.06	-3.510960978	pronoun + verb + to + intrans-verb, nothing after	**5.43.g She tried to leave.**
-0.40	-5.532150277	pronoun + verb + to + tensed-verb	**5.45.* She tried to left.**
-1.02	-5.532150277	pronoun + verb + to + modal + intrans-verb	**5.47.* She tried to may leave.**
-0.54	-5.532150277	pronoun + verb + to + do + intrans-verb	**5.49.* She tried to do leave.**
0.27	-5.532150277	name + verb + to + verb-intrans + and + name + verb + to + adverb	**5.50.g Casey wanted to sleep and Marcy tried to as well.**
-0.60	-5.532150277	name + verb + to + verb-intrans + and + name + verb + to + do	**5.51.* Casey wanted to sleep and Marcy tried to do.**
0.97	-5.532150277	pronoun-subj + modal + aux-have + aux-be + tran-verb-VBG + simple-object	**5.77.g I might have been eating dinner.**
-0.76	-5.532150277	pronoun-subj + aux-have + modal + be + trans-verb-VBG + simple-object	**5.81.* I have might be eating dinner.**
-0.68	-5.532150277	pronoun-subj + aux-have + verb-part + to + aux-have + verb-part2, + and + aux-have + verb-part2 + pronoun + aux-do	**5.84.* I'd planned to have finished, and have finished I did.**
Acceptability	($\log_{10}$) Frequency	Salient Structure	Instantiation of Structure (Utterance)
-0.33	-5.532150277	pronoun-subj + aux-have + verb-part + to +	**5.84.g I'd planned to have finished, and finished I have.**

		aux-have + verb-part2, + and + verb-part2 + pronoun + aux-have	
0.93	-5.532150277	name + aux-has + aux-be-participle + VBG-verb + PP[=prep+name], nothing after in clause	**5.92.g Jason has been arguing with Noel.**
-0.69	-5.532150277	name + aux-be+aux-have + participle + PP=[prep + name]	**5.93.* Jason is having argued with Noel.**
-1.19	-5.532150277	name + neg + past-tense-verb-trasn + simple object	**5.133.* Ryan not flew the airplane.**
1.13	-5.532150277	name + do-aux + neg + trans-verb + simple object	**5.135.g Ryan did not fly the airplane.**
1.06	-4.687052237	name + aux-has + adverb + trans-verb-participle + simple object	**5.139.g Ryan has never flown an airplane.**
0.47	-5.532150277	name + adv + has-aux + trans-verb-participle + simple object	**5.140.* Ryan never has flown an airplane.**
0.91	-5.532150277	name + aux-has-n't + past-participle-verb-intransitive	**5.144.g Jason hasn't arrived.**
-0.89	-5.532150277	name + not + past-tense-verb-intransitive	**5.145.* Jason not arrived.**
1.00	-5.532150277	name + didn't + non-finite-verb-intransitive	**5.146.g Jason didn't arrive.**
0.05	-5.532150277	name + didn't + past-tense-verb-intransitive	**5.147.* Jason didn't arrived.**
0.66	-5.532150277	subj=[quant + det + plural noun] + aux-had + verb-intrans-past-participle	**6.5.g All the horses had escaped.**
0.63	-5.532150277	subj=[dt + plural noun] + have-aux + quantifier + verb-intrans-past-participle	**6.7.g The horses had all escaped.**
-0.53	-5.532150277	subj=[bare plural noun] + have-aux + quantifier-adj + be-aux + trans-verb-passive	**6.9-.* Horses have most been domesticated.**
Acceptability	($\log_{10}$) Frequency	Salient Structure	Instantiation of Structure (Utterance)
1.11	-5.532150277	subj=[quant + noun] + have-aux + be-aux + trans-verb-passive	**6.9.g Most horses have been domesticated.**
0.80	-4.833180273	subj=[nom pro] + have-	**6.38.g She has kissed her.**

		aux + trans-verb-past-participle + obj=[acc pro]	
-1.04	-5.532150277	subj=[acc pro] + have-aux + trans-verb-past-participle + obj=[acc pro]	**6.39.* Her has kissed her.**
-1.03	-5.532150277	subj=[nom pro] + have-aux + trans-verb-past-participle + obj=[nom pro]	**6.40.* She has kissed she.**
-0.72	-5.532150277	expletive there + verb + obj=[nom pronoun] + PP-simple-obj	**6.45a.* There was he in the garden.**
-0.81	-5.532150277	expletive there + verb + obj=[acc pronoun] + PP-simple-obj	**6.45b.* There was him in the garden.**
1.10	-3.776275421	expletive there + verb + simple NP + PP-simple-obj	**6.45c.g There was a man in the garden.**
-0.97	-5.532150277	object=[accusative-pronoun] + have-aux +subject=[nom-pronoun] + past-participle-transitive-verb	**6.58.* Him has he known.**
0.62	-4.833180273	pronoun-subject + have-aux + past-participle-transitive + pronoun-object=not-refl	**6.58.g He has known him.**
-0.47	-5.532150277	simple subject + be-verb + past-tense-transitive-verb	**6.93.* The clothes were stole.**
0.76	-4.356059018	simple subject + be-verb + passive-transitive-verb	**6.93.g The clothes were stolen.**
0.89	-5.055029022	simple subject + be + passive verb + PP-doer=[by + name]	**6.98.g The boy was killed by Stan.**
-0.53	-5.532150277	simple subject + intransitive verb + instrumental PP=[doer-by+name]	**6.99.* The boy arrived by Stan.**
-1.00	-5.532150277	expl-there + intransitive verb + PP=[by+name]	**6.100.* There arrived by Stan.**
-0.57	-5.532150277	expl-there + be + passive + simple object + PP=[by + simple object]	**6.102.* There were killed three men by the assassin.**
Acceptability	($\log_{10}$) Frequency	Salient Structure	Instantiation of Structure (Utterance)
1.06	-4.490757592	simple subject + be + passive verb + PP=[by + simple object]	**6.102.g Three men were killed by the assassin.**

0.06	-5.532150277	simple subject + adverb + modal + transitive verb + simple object	**6.106.* Elliot quickly may free the animals.**
0.83	-4.301701356	simple subject + modal + adverb + transitive verb + simple object	**6.107.g Elliot may quickly free the animals.**
0.39	-5.532150277	simple subject + modal + adverb + have + perfective transitive verb + simple object	**6.108.g Elliot could quickly have freed the animals.**
-0.63	-5.532150277	simple subject + transitive tensed verb (not be) + adverb + simple object	**6.112.* Garry failed often calculus exams.**
0.99	-3.878937763	simple subject + adverb + transitive tensed verb (not be) + simple object	**6.112.g Garry often failed calculus exams.**
1.04	-4.100786513	NP = [determiner + plural count noun] + verb + PP, nothing after	**7.03.g The letters are on the table.**
-0.96	-5.532150277	NP = [plural count noun + determiner] + verb + PP, nothing after	**7.04.* Letters the are on the table.**
1.07	-4.069752279	NP = [plural count noun with no determiner] + verb + PP, nothing after	**7.06.g Letters are on the table.**
-0.09	-4.356059018	NP = [singular count noun with no determiner] + verb + PP, nothing after	**7.07.* Letter is on the table.**
-0.30	-5.532150277	NP = [determiner + demonstrative + simple noun] + transitive verb + NP = [det + simple noun], nothing after	**7.30.* The this man needs a taxi.**
1.20	-4.490757592	NP = [demonstrative + simple noun] + transitive verb + NP = [det + simple noun], nothing after	**7.30.g This man needs a taxi.**
1.07	-4.356059018	NP = [proper name possessive + simple noun] + verb + adjective	**7.52.g Evan's idea is brilliant.**
Acceptability	($\log_{10}$) Frequency	Salient Structure	Instantiation of Structure (Utterance)
-0.90	-5.532150277	complex NP = [possessive + det + simple noun] + verb + adjective	**7.54.* Evan's the idea is brilliant.**
-1.02	-5.532150277	NP = [nom pro + noun + simple-noun] + verb +	**7.89.* He analysis her was flawed.**

		participle	
0.39	-5.532150277	NP = [possessive + noun + PP-of-simple-noun] + verb + participle	**7.90.g His analysis of her was flawed.**
-0.70	-5.532150277	complex NP = [simple noun + PP-of-possessive-basic form] + verb + PP with simple object	**7.103.* A book of my is on the desk.**
0.58	-5.532150277	complex NP = [simple noun + PP-of-possessive] + verb + PP with simple object	**7.104.g A book of mine is on the desk.**
0.13	-5.532150277	complex subj = [possessive + noun + PP-of with possessive NP object] + verb + participle, nothing after	**7.105.* The therapist's analysis of Morticia's was flawed.**
0.83	-5.532150277	complex subj = [possessive + noun + PP-of with simple NP] + verb + participle, nothing after	**7.105.g The therapist's analysis of Morticia was flawed.**
-0.51	-5.532150277	complex subject = [wh-obj + simple subject + transitive verb + comp-that] + verb + complex object = [simple subject + verb + adjective/participle]	**8.03.* What she thought that was the poison was neutralized.**
0.18	-5.532150277	complex subject = [wh-obj + simple subject + transitive verb] + verb + complex object = [comp-that + simple subject + verb + adjective/participle]	**8.03.g What she thought was that the poison was neutralized.**
1.07	-5.532150277	quantified subject + verb + complex object = [comp-that + simple subj + verb + adjective]	**8.05.g Everyone claimed that the wedding was beautiful.**
-0.34	-5.532150277	complex subj = [comp-that + simple subj + verb + adjective] + verb + transitive verb-passive + PP-by-phrase	**8.06.? That the wedding was beautiful was claimed by everyone.**
Acceptability	($\log_{10}$) Frequency	Salient Structure	Instantiation of Structure (Utterance)
1.13	-5.532150277	simple subject + wonder-verb + complex object = [comp-whether + simple subject + verb	**8.19.g Jason wondered whether the potion was ready.**

		+ adjective]	
0.19	-5.532150277	simple subject + wonder-verb + complex obj = [comp-wonder + comp-that + simple subject + verb + adjective]	**8.21.* Jason wondered whether that the potion was ready.**
0.31	-5.532150277	complex subject = [wh-obj + simple subject + wonder-verb]+ verb + complex object = [comp-whether + simple subject + verb + adjective]	**8.23.g What Jason wondered was whether the potion was ready.**
-0.57	-5.532150277	complex subj = [wh-obj + simple subj + wonder-verb + whether-comp] + verb + simple subj + adjective	**8.24.* What Jason wondered whether was the potion ready.**
-0.34	-5.532150277	simple subject + wonder-verb + complex object = [comp-that + simple subject + verb + adjective]	**8.29.* Jason wondered that the potion was ready.**
1.13	-5.532150277	simple subject + wonder-verb + complex object = [comp-whether + simple subject + verb + adjective]	**8.29.g Jason wondered whether the potion was ready.**
-0.10	-5.532150277	complex subject = [comp-that + simple subject + verb + adjective]+ transitive verb + simple object	**8.56.g That the answer is obvious upset Helen.**
-0.65	-5.532150277	complex subject = [comp that + complex subject =[comp that + simple subject + verb + adjective] + verb + adjective]+ transitive verb + simple object	**8.57.* That that the world is round is obvious upset Helen.**
-0.54	-5.532150277	complex subject = [comp that + complex subject = [whether + simple subj + verb + adjective] + verb + adjective] + transitive verb + simple object	**8.58.* That whether the world is round is unknown upset Helen.**
Acceptability	($\log_{10}$) Frequency	Salient Structure	Instantiation of Structure (Utterance)
-0.60	-5.532150277	complex subj = [comp that + simple subj +	**8.64.* That Jason had arrived was obvious annoyed Mandy.**

		intransitive verb] + verb + adjective + transitive verb + simple object	
-0.66	-5.532150277	simple subj + transitive verb + compl that + complex obj = [complex subj =[comp-that + simple subj + intransitive verb] + transitive verb + simple object]	**8.65.* I said that that Jason had arrived annoyed Mandy.**
0.24	-5.532150277	expletive there + intransitive verb + simple noun + PP, nothing after	**8.70.g There arrived a new actor on the set.**
-0.73	-5.532150277	simple subject + intransitive verb + object + PP, nothing else after.	**8.71.* The director arrived a new actor on the set.**
0.99	-5.532150277	simple subj + verb + complex object = infinitive + transitive verb + simple object	**8.74.g Laura tried to bathe her children.**
-1.02	-5.532150277	simple subj + verb + complex object = same simple subject + infinitive + transitive verb + simple object	**8.76.* Laura tried Laura to bathe her children.**
-0.86	-5.532150277	simple subj + verb + complex object = simple subject + infinitive + transitive verb + simple object	**8.77.* Laura tried the babysitter to bathe her children.**
-0.81	-5.532150277	simple subj + transitive verb + null object + infinitive verb with adjective following	**8.92.* We believed to be omnipotent.**
0.61	-5.532150277	quantified subject + verb + complex object = infinitive + verb, and nothing after	**8.93.g No one expected to win.**
0.86	-4.833180273	simple subject + transitive verb + complex obj = for + accusative pronoun subject + infinitive with transitive verb + simple object	**8.102.g Brian intended for him to learn magic.**
Acceptability	($\log_{10}$) Frequency	Salient Structure	Instantiation of Structure (Utterance)
-0.69	-5.532150277	simple subject + transitive verb +	**8.104.* Brian intended for to learn magic.**

		complex obj = for + infinitive with transitive verb + simple object	
-0.57	-5.532150277	for + infinitive with transitive verb = complex subject + verb + simple object	**8.105.* For to do that would be a mistake.**
0.53	-5.532150277	for + simple NP + infinitive with transitive verb = complex subject + verb + simple object	**8.105.g For him to do that would be a mistake.**
0.55	-5.532150277	simple subj + transitive verb + simple object + infinitive verb with adjective following	**8.120.g We believed him to be omnipotent.**
0.48	-5.532150277	simple WH-obj + simple subject + transitive verb + connective "to be" + "for" PP + infinitive with transitive verb and simple object	**8.131.g What Brian intended was for him to learn magic.**
0.31	-5.532150277	simple WH-obj + simple subject + transitive verb + connective "to be" + infinitive with transitive verb + simple object	**8.132.g What Brian tried was to learn magic.**
-0.65	-5.532150277	complex WH subject (WH-obj + simple subj + transitive verb) + tensed verb + non-finite clause (with accusative subject + infinitive verb + simple adjective)	**8.133. *What Brian believed was him to be omnipotent.**
-0.93	-5.532150277	Simple subject + seem + comp that + present tense verb + adjective, nothing else in embedded clause	**8.150. *Melissa seems that is happy.**
0.83	-5.532150277	Expletive it + seem + comp that + simple subject + tensed verb + simple adjective	**8.151.g It seems that Melissa is happy.**
1.00	-5.055029022	simple subject + seem + infinitive verb + adjective, nothing else in clause	**8.152.g Melissa seems to be happy.**
Acceptability	($\log_{10}$) Frequency	Salient Structure	Instantiation of Structure (Utterance)
-0.75	-5.532150277	subject + stink + infinitive be + adjective	**8.167.*Zeus stinks to be omnipotent.**
0.69	-5.532150277	expetive there + seem + infinitive + NP + simple	**8.168.g There seems to be a man in the garden.**

		PP	
-0.27	-5.532150277	expetive there + seem + NP + infinitive + PP	**8.176. *There seems a man to be in the garden.**
0.25	-5.532150277	one-word subject + expect + expetive there + infinitive transitive verb with object, nothing else in clause	**8.148.g I expected there to be a problem.**
-0.48	-5.532150277	one-word subject + persuade + expetive there + infinitive transitive verb with object, nothing else in clause	**8.185. *I persuaded there to be a problem.**
1.12	-3.592631024	Who + aux + NP-SBJ (single word) + transitive verb, nothing else in clause	**9.04.g Who did Nancy poison?**
-1.01	-5.532150277	NP topic + aux + NP subject + transitive verb, nothing else in clause	**9.12.* Someone did Nancy poison.**
-0.64	-5.532150277	where + nn(s) under WHNP + subject + transitive verb, nothing else in clause	**9.25. *Where place are you living?**
1.03	-5.055029022	which + nn(s) under WHNP + subject + transitive verb, nothing else in clause	**9.28.g Which poem did Harry recite?**
-0.35	-5.532150277	which + dt + noun under WHNP + subj + transitive verb	**9.32. *Which the poem did Harry recite?**
0.18	-5.532150277	subj + wonder verb + embedded y/n with aux + subj + intransitive verb + adverb, nothing after in clause	**9.83. *I wondered could we leave early.**
1.03	-5.532150277	subj + wonder verb + tensed if clause with subj, aux, intrans verb, adverb, nothing else in clause	**9.83.g. I wondered if we could leave early.**
Acceptability	($\log_{10}$) Frequency	Salient Structure	Instantiation of Structure (Utterance)
0.89	-5.532150277	subj + wonder verb + tensed embedded with wh-obj fronted + subj + transitive verb, nothing after in clause	**9.84.g. I wondered who Nancy poisoned.**
-0.80	-5.532150277	subj + think verb + tensed embedded with	**9.105.* Jason thinks who Nancy poisoned.**

		wh-obj fronted + subj + transitive verb	
-0.54	-5.532150277	WH-obj + Wh-subj in situ + transitive verb	**9.120. *Who did who poison?**
0.11	-5.532150277	WH-subj + transitive verb + WH-obj in situ	**9.120.g. Who poisoned who?**
0.01	-5.532150277	WH + subj + ditrans verb + obj trace + pp with WH in situ	**9.122.g. Who did Anna introduce to whom?**
-0.21	-5.532150277	WH + subj + ditrans verb + wh-obj + pp with trace	**9.123. *Who did Anna introduce who to?**
0.40	-5.532150277	WDT-subj + transitive verb + WDT-obj	**9.124.g. Which poet wrote which poem?**
-0.10	-5.055029022	WDT-obj + WDT-subj + transitive verb, nothing after in clause	**9.125.g. Which poem did which poet write?**
0.27	-5.532150277	Subj + verb + embedded clause with wh-subj and wh-obj + transitive verb	**10.55.g I asked who poisoned who.**
-0.91	-5.532150277	Subj + verb + embedded clause with wh-obj first then wh-subj + transitive verb	**10.56. *I asked who who poisoned.**
-0.90	-5.532150277	WH + subj + verb + embedded wh-subj + trans verb + trace of object	**10.58. *Who did you ask who poisoned?**
0.60	-4.134210268	subject + verb + np containing that-clause with subject and transitive verb and object (without CP in it), nothing after in clause	**10.69.g I believed the claim that Philip would visit the city of Athens.**
-0.43	-5.055029022	wdt + noun wh-phrase + subj + verb + complex noun with that-clause that has subj and transitive verb, nothing else after in clause	**10.70. *Which city did you believe the claim that Philip would visit?**
Acceptability	($\log_{10}$) Frequency	Salient Structure	Instantiation of Structure (Utterance)
1.07	-5.532150277	subj + verb + PP with complex noun, which has possessive NP spec and PP (not of) complement with overt object, nothing after in clause	**10.71.g. Peter listened to Darren's speech about investment banks.**
-0.45	-5.532150277	[main clause] WH + subj	**10.72. *What did Peter listen to**

		+ verb + PP with complex noun object, which has possessive NP and PP complement	**Darren's speech about?**
0.95	-5.532150277	subj + aux + VBN/JJ + PP with complex noun with possessive NP and with embedded PP with of, nothing after	**10.73.g Penny was interested in Philip's description of geometry class.**
-0.24	-5.532150277	[main clause] WH + subj + verb + adjp + PP with complex noun with embedded PP (with of)	**10.74. *What was Penny interested in Philip's description of?**
-0.30	-5.532150277	[main clause] WH + subj + verb + PP with complex noun with embedded PP (not of) and demonstrative determiner	**10.83. *What did Peter listen to those speeches about?**
0.06	-5.532150277	that-clause subject with object extraction + adjective pred	**10.90.g. That Peter loved Amber was obvious.**
1.02	-5.532150277	Expletive + adjectove pred + that-clause with object extraction	**10.91.g. It was obvious that Peter loved Amber.**
-0.39	-5.532150277	WH + adjective + that-clause subject with object extraction	**10.92.g Who was it obvious that Peter loved?**
-1.04	-5.532150277	WH + that-clause subject with object extraction + adj predicate	**10.93. *Who was that Peter loved obvious?**
0.05	-5.532150277	that-clause subject with object extraction + seem + non-finite clause + PP	**10.94.g That Peter loved Amber seemed to be known by everybody.**
-0.97	-5.532150277	WH + that-clause subject with object extraction + seem + non-finite clause + PP	**10.95. *Who did that Peter loved seem to be known by everybody?**
0.97	-4.833180273	subject with PP (not "of") [not conjunct NP] + additional PP	**10.107a.g A program about Elephants is on channel 4 tonight.**
Acceptability	($\log_{10}$) Frequency	Salient Structure	Instantiation of Structure (Utterance)
1.02	-4.833180273	expletive t?here + subject with PP (not "of") + additional PP	**10.108a.g. There is a program about Elephants on channel 4 tonight.**
-0.04	-5.532150277	WH + expletive there + subject with PP extraction (prep overt) + PP	**10.108b.g. What is there a program about on channel 4 tonight?**
0.58	-5.055029022	main verb takes only	**10.116.g. I worried after the**

		"after" embedded clause with transitive verb + object + locations PP.	**lawyer forgot his briefcase at the office.**
0.73	-4.687052237	main verb takes only "because" embedded clause with transitive verb + locations PP	**10.117.g. I worried because the lawyer forgot his briefcase at the office.**
0.18	-4.253396676	main verb takes "if" embedded clause with transitive verb + object + locations PP.	**10.118.g. I worry if the lawyer forgets his briefcase at the office.**
-0.79	-5.532150277	main verb takes "after" embedded clause with transitive verb + locational PP	**10.119. *What did you worry after the lawyer forgot at the office?**
-0.73	-5.532150277	main verb takes "because" embedded clause with transitive verb + locational PP	**10.120. *What do you worry because the lawyer forgot at the office?**
-0.61	-5.532150277	main verb takes "if" embedded clause with transitive verb + locational PP	**10.121. *What do you worry if the lawyer forgets at the office ?**