

INVESTIGATING PROSODIC BOUNDARIES

A Thesis

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Master of Arts

by

Serena Crivellaro

August 2008

© 2008 Serena Crivellaro

ABSTRACT

This thesis relates the results of a series of experiments testing speakers' interpretation of ambiguous sentences in which prosodic cues at two relevant locations have been systematically manipulated. The goal was to develop a model to predict sentence's interpretations based on information about the relative strength of the two boundaries.

The model we present builds on Carlson, Clifton and Frazier's (2001) Informative Boundary Hypothesis, which proposes that attachments are derived based on the relative strength of the two phonetic boundaries, and introduces room for variability based on the effects of Overall Bias and Conditional Bias on the distribution of results. Overall Bias is a bias introduced into the distribution of responses that affects all tokens, shifting the probabilities of each interpretation for every token by the same amount. Conditional Bias is instead defined as a processing-related bias, that reduces the probability of one or the other interpretations only for boundaries larger than a certain amount at a specific location in the sentence. Extra time at this boundary allows the speaker to process the sentence, which makes one or the other readings unlikely.

To test the model, we generated a set of sentences with equally spaced boundaries at each of two locations in the sentence that would be consistent with high or low attachment interpretations, thus generating a grid-like structure that can be used as the base for a map of the sentences' interpretations. The tokens were generated using an automated script from a single base file which in turn had been produced by a speech synthesizer, to minimize the amount of variability across tokens. Subjects were then asked to select an interpretation and a confidence rating for each token, which were then combined to result in a weighted response variable.

The distribution of the responses, and the point at which interpretations change, can be plotted for each of the items. The predictions about Conditional Bias made on the grounds of sentence processing were borne out for almost all the structures. Some items also showed the effects of Overall Bias, and the remaining part of the cases can be explained by introducing the effects of the experimental task, which can trigger earlier processing points under very specific circumstances. Furthermore, it appears that the amount of boundary which triggers a Conditional Bias effect is consistent across items and structures, suggesting that a systematic threshold is necessary for processing to take place, which should be investigated further in future research.

BIOGRAPHICAL SKETCH

Serena was born near Verona, in Italy, in 1985. Despite technically being Italian, she spent most of her life abroad, which fostered her interest in languages and linguistics. She attended Yale College for her undergraduate studies and received a B.A. in Linguistics in May 2006, with a thesis entitled "The Syntax of XIII Comuni Cimbrian: Contact-induced change in an Endangered Language". She then proceeded to Cornell University for her graduate education in Linguistics, and is fulfilling her Master of Arts' Degree requirements with this thesis.

To Jacopo

*Because nothing motivates me more
than wanting to be the best possible role model
for my most inspiring scholar-to-be*

ACKNOWLEDGMENTS

This thesis would not have been possible without the help and support of many people in all areas of my life. I would especially like to thank Michael Wagner and Draga Zec, my advisors, for giving me guidance throughout all stages of the project, from the initial idea development, through the experiment setup, and to the data analysis and final presentation. I also want to thank Sue Hertz for offering Phonetics help and for giving me access to a Demo version of the Eti-Eloquence synthesizer to produce the stimuli used in this set of experiments. And I am also grateful to Ted Gibson and Mara Breen, for helping me develop the initial idea that led to this thesis during a summer opportunity at MIT's TedLab.

This thesis would have never seen the light without the help of Johanna Brugman, instrumental in teaching me how to script in Praat and always willing to help debug or simplify problematic loops, and I also would like to thank Eric Evans, for offering great technical and Lab support and being so patient with me as I took over the Plab to run subjects.

Furthermore, I would like to thank Peggy Renwick, Adam Cooper, Ed Cormany, Masayuki Gibson, Steven Ikier, and the rest of the CLC for always being there to bounce ideas off of, complain to, get grammaticality judgments from, order late-night takeout with, and for generally trying to keep me sane. And lastly, but certainly not least, I would like to thank my family for always being there to hold my hand, keep me focused, and remind me that it was time to take a break or go to bed.

I couldn't have done it without you all: Grazie!

TABLE OF CONTENTS

BIOGRAPHICAL SKETCH.....	iii
ACKNOWLEDGMENTS.....	v
TABLE OF CONTENTS.....	vi
LIST OF FIGURES.....	xi
PROSODIC BOUNDARIES.....	1
0.0 Introduction.....	1
0.1 Previous Work in Prosodic Boundary Perception.....	1
0.1.1 Perception Experiments.....	1
0.1.2 Production Studies.....	5
0.1.3 Communication Experiments.....	7
0.2 Our Hypothesis.....	8
0.2.1 The Experiment Setup.....	8
0.2.2 Overall Bias.....	9
0.2.3 Conditional Bias.....	13
0.2.4 Predictions of the Model.....	16
STUDY ONE: PHONETIC CUES.....	19
1.0 Introduction.....	19
1.1 Method.....	19
1.1.1 Stimuli.....	20
1.1.2 Subjects.....	21
1.1.3 Experimental Setup.....	21
1.1.4 Variables.....	22
1.2 Results.....	23
1.3 Discussion.....	24
STUDY TWO: SIMPLE CONJUNCTIONS.....	26

2.0 Introduction	26
2.1 Method.....	26
2.1.1 Stimuli	26
2.1.2 Subjects.....	28
2.1.3 Experiment Setup	28
2.1.4 Conditional Bias Predictions	30
2.2 Results	31
2.2.1 Algebraic Formulas	31
2.2.2 Suspects B and C and D	34
2.2.3 Rose and Steve and Kim.....	37
2.2.4 Eve or Jude and Sue	41
2.3 Discussion.....	43
STUDY THREE: MODIFIED CONJUNCTIONS	46
3.0 Introduction	46
3.1 Method.....	46
3.1.1 Stimuli	46
3.1.2 Subjects.....	47
3.1.3 Experiment Setup	48
3.1.4 Conditional Bias Predictions	49
3.2 Results	49
3.2.1 Dancers and Skaters	49
3.2.2 Farmers and Workers	51
3.2.3 Chefs and Wine-Tasters	53
3.3 Discussion.....	54
STUDY FOUR: PARTICLE VERBS	56
4.0 Introduction	56

4.1 Method.....	56
4.1.1 Stimuli	56
4.1.2 Subjects.....	57
4.1.3 Experimental Setup	57
4.1.4 Conditional Bias Predictions	58
4.2 Results	59
4.2.1 Check in.....	59
4.2.2 Drop off	62
4.2.3 Win Over	64
4.2.4 Wear Down.....	68
4.2.5 Look Up.....	69
4.3 Discussion.....	71
STUDY FIVE: PREPOSITIONAL PHRASES	76
5.0 Introduction	76
5.1 Method.....	76
5.1.1 Stimuli	76
5.1.2 Subjects.....	77
5.1.3 Experiment Setup	77
5.1.4 Conditional Bias Predictions	79
5.2 Results	80
5.2.1 Teddy bears	80
5.2.2 Rottweilers.....	84
5.2.3 Worried Expressions	86
5.2.4 Offending Bows.....	87
5.2.5 Cannons	89
5.2.6 Attack Plans.....	91

5.3 Discussion.....	93
STUDY SIX: RELATIVE CLAUSES.....	95
6.0 Introduction	95
6.1 Method.....	95
6.1.1 Stimuli	95
6.1.2 Subjects.....	96
6.1.3 Experimental Setup	96
6.1.4 Conditional Bias Predictions	96
6.2 Results	97
6.2.1 The Daughter of the Colonel	97
6.2.2 The Killer of the Journalist.....	99
6.3 Discussion.....	101
STUDY SEVEN: MIDDLE ATTACHMENTS	102
7.0 Introduction	102
7.1 Method.....	102
7.1.1 Stimuli	102
7.1.2 Subjects.....	103
7.1.3 Experimental Setup	103
7.1.4 Conditional Bias Predictions	104
7.1.5 Peculiarities of the Structure	105
7.2 Results	106
7.2.1 In the woods	107
7.2.2 Clearly	108
7.2.3 Gradually.....	111
7.3 Results	112
GENERAL DISCUSSION.....	114

8.0 Summary of Results	114
8.1 Overall Bias	118
8.2 Conditional Bias	119
8.2.1 The Predictions of Our Processing Model.....	120
8.2.2 Phonetic Properties of Conditional Bias.....	121
8.2.3 Variability in the Domain of Conditional Bias.....	123
8.3 The Interaction of Overall and Conditional Bias	125
8.4 Concluding Remarks	127
REFERENCES	129

LIST OF FIGURES

Figure 1: The Effect of Overall Bias on the Distribution Plot	11
Figure 2: The Effect of Overall Bias on the Difference Variable	12
Figure 3: An Example of Conditional Bias in Action	14
Figure 4: The Effect of Conditional Bias on the Difference Variable	15
Figure 5: Predicted Contour Lines	16
Figure 6: More Predicted Contour Lines.....	17
Figure 7: Response Distribution for Study One Items	24
Figure 8: Comparing Maximum Mean Response Displacements	25
Figure 9: Distribution Plot for item "B plus C times D"	32
Figure 10: Difference Bar-chart for item "B plus C times D"	33
Figure 11: Distribution Plot for item "B and C and D" (short)	35
Figure 12: Difference Bar-chart for item "B and C and D" (short).....	35
Figure 13: Distribution Plot for item "B and C and D" (long)	36
Figure 14: Difference Bar-chart for item "B and C and D" (long).....	37
Figure 15: Distribution Plot for item "Rose and Steve and Kim" (short).....	38
Figure 16: Difference Bar-chart for item "Rose and Steve and Kim" (short)	39
Figure 17: Distribution Plot for item "Rose and Steve and Kim" (long)	40
Figure 18: Difference Bar-chart for item "Rose and Steve and Kim" (long).....	41
Figure 19: Distribution Plot for item "Eve or Jude and Sue"	42
Figure 20: Difference Bar-chart for item "Eve or Jude and Sue"	42
Figure 21: Predicted Contour Line for First Boundary Conditional Bias	44
Figure 22: Distribution Plot for item "Dancers and Skaters"	50
Figure 23: Difference Bar-chart for item "Dancers and Skaters"	51
Figure 24: Distribution Plot for item "Farmers and Workers"	52

Figure 25: Difference Bar-chart for item "Farmers and Workers"	52
Figure 26: Distribution Plot for item "Chefs and Wine-tasters"	53
Figure 27: Difference Bar-chart for item "Chefs and Wine-tasters"	54
Figure 28: Distribution Plot for item "Check In" (short).....	60
Figure 29: Difference Bar-chart for Item "Check In" (short).....	60
Figure 30: Distribution Plot for item "Check In" (long)	61
Figure 31: Difference Bar-chart for item "Check In" (long)	62
Figure 32: Distribution Plot for Item "Drop Off"	63
Figure 33: Difference Bar-chart for item "Drop Off"	64
Figure 34: Distribution Plot for Item "Win Over" (short)	65
Figure 35: Difference Bar-chart for Item "Win Over" (short).....	66
Figure 36: Distribution Plot for Item "Win Over" (long).....	66
Figure 37: Difference Bar-chart for Item "Win Over" (long)	67
Figure 38: Distribution Plot for Item "Wear Down"	68
Figure 39: Difference Bar-chart for Item "Wear Down"	69
Figure 40: Distribution Plot for Item "Look Up"	70
Figure 41: Difference Bar-chart for Item "Look Up"	71
Figure 42: Comparative Difference Bar-charts for Particle Verb items.....	72
Figure 43: Predicted Interaction Effect of Strong Late Break Overall Bias and First Boundary Conditional Bias	74
Figure 44: Distribution Plot for "With the Teddy Bear" (short).....	80
Figure 45: Distribution Plot for Item "With the Teddy Bear" (long, 7)	82
Figure 46: Difference Bar-chart for Item "With the Teddy Bear" (long, 7).....	82
Figure 47: Distribution Plot for Item "With the Teddy Bear" (long, 4)	84
Figure 48: Difference Bar-chart for Item "With the Teddy Bear" (long, 4).....	84
Figure 49: Distribution Plot for Item "With the Rottweiler"	85

Figure 50: Difference Bar-chart for Item "With the Rottweiler"	85
Figure 51: Distribution Plot for Item "With a Worried Expression"	86
Figure 52: Difference Bar-chart for the Item "With a Worried Expression"	87
Figure 53: Distribution Plot for Item "With a Bow"	88
Figure 54: Difference Bar-chart for Item "With a Bow"	89
Figure 55: Distribution Plot for Item "With the Cannon"	90
Figure 56: Difference Bar-chart for Item "With the Cannon"	91
Figure 57: Distribution Plot for Item "With the Attack Plan"	92
Figure 58: Difference Bar-chart for Item "With the Attack Plan"	92
Figure 59: Distribution Plot for Item "Daughter of the Colonel"	98
Figure 60: Difference Bar-chart for Item "Daughter of the Colonel"	99
Figure 61: Distribution Plot for Item "Killer of the Journalist"	100
Figure 62: Difference Bar-chart for Item "Killer of the Journalist"	101
Figure 63: A Possible Graphical Representation of Topicalization-Bias.....	105
Figure 64: Distribution Plot for Item "In the Woods"	107
Figure 65: Difference Bar-chart for Item "In the Woods"	108
Figure 66: Distribution Plot for Item "Clearly"	109
Figure 67: Difference Bar-chart for Item "clearly"	110
Figure 68: Distribution Plot for Item "Gradually"	111
Figure 69: Difference Bar-chart for Item "Gradually"	112
Figure 70: Summary of Overall and Conditional Bias Results	114
Figure 71: Reciprocally Obscuring Effects of the Interaction of Overall and Conditional Bias	126

PROSODIC BOUNDARIES

0.0 Introduction

The studies presented in this thesis are designed to test the relative import of prosodic boundaries in the disambiguation of syntactically ambiguous sentences. Clifton, Carlson and Frazier (2001), in their Informative Boundary Hypothesis, proposed that all boundaries (normally no more than two) within a relevant domain are taken into account in reconstructing the meaning of the ambiguous sentence, while previous accounts assumed that the single prosodic boundary corresponding to the syntactic break would key the listener into the intended meaning. This thesis is a collection of studies which, through controlled manipulations of boundary sizes, aims to determine whether, which, and how much, prosodic boundaries at different locations in different syntactic structures contribute to meaning disambiguation.

0.1 Previous Work in Prosodic Boundary Perception

In this section, we briefly review previous studies of prosodic boundaries in ambiguous syntactic structures, their findings and conclusions, and how we hope to expand on their conclusions in the present investigation.

0.1.1 Perception Experiments

In her 1973 paper “Phonetic Disambiguation of Syntactic Ambiguity”, Ilse Lehiste tested speakers’ comprehension of recordings of ten grammatically (syntactically or lexically) ambiguous sentences, recorded from speakers in three states: before they were made aware of the ambiguity, and then after they were debriefed with both possible meanings in mind. Subjects were asked to pick the intended meaning of each of the three recordings per sentence, and the most accurately identified sentences were then examined to determine the common successful

disambiguating cues. Lehiste found that timing (specifically pauses), “drawls” (which we now call pre-boundary lengthening), and the amount of laryngealization preceding a boundary were more effective cues than the use of F0 (pitch), which was used mostly for non-syntactic disambiguation.

In the same paper, she also tested the then-current hypothesis relating syntactic and prosodic structure and interpretation (Lieberman 1967), namely that prosodic information can only be used to disambiguate surface structure ambiguities, and not deep-structure, or label-only ambiguities.

In 1976, Lehiste, Olive and Streeter further investigated the same hypothesis and found once again that lexically ambiguous sentences¹ were not reliably differentiated through boundary placement and strength, unlike syntactically ambiguous sentences such as the ones described above. However, in this iteration of the experiment, they began to use more sophisticated experimental protocols, and instead of recording a different sound file for each item, they recorded naïve speakers producing the sentences once, and obtained meaning ratings for each sound file. The most ambiguous (neutral-sounding) token of each set was then selected to serve as the basis for future manipulation. In their manipulations, they flattened the pitch of all the sentences, and then systematically manipulated the duration of interstress intervals, and presented subjects with these new tokens for evaluation.

Lehiste’s phonetic findings were expanded upon by Lynn Streeter in 1978, with two experiments investigating the relative importance of Duration, Intensity, and Intonation cues in sentence interpretation. Using formulas of the type “A + E * O”, recorded with both bracketing structures by two different speakers, Streeter

¹ Such as “German teachers visit Greensboro” (Lehiste 1976)

manipulated the sentences varying each cue independently, replacing the values for one bracketing structure with that of the other, and then asked subjects to indicate the intended meaning. She found that Duration and Intonation had significant, additive, effects on meaning change, and that Intensity only reinforced existing cues but didn't change the interpretation of the sentence.

Advances in technology now allow the use of synthetic speech, rather than human speech, as the base sound file from which the experimental items are derived, which allows for even more control over the rate of speech and phonetic detail of the sentence, to ensure that even tokens across different items are comparable. For the sound tokens in this experiment, we will be using sound files generated by a Demo version of the ETI-Eloquence “Elocutor” synthesizer, a formant-based synthesizer: although sound files produced by an intelligent synthesizer do carry prosodic cues, unlike human speech, these are also systematic and constant across items, so after some experimentation it is possible to calculate which operations were applied to produce this prosody, and if necessary, undo them to produce a non-prosodically-biased version of the sentence. This prosodically neutral sentence can then be automatically manipulated, in our case via a series of Praat scripts that utilize the program's PSOLA function, to generate a number of tokens that differ only in the size of the prosodic boundaries at the relevant locations.

In the following years attention shifted to different types of experiments (discussed in sections 0.1.2 and 0.1.3), but our model is built mostly on the findings presented in Carlson, Clifton and Frazier's 2001 paper entitled “Prosodic Boundaries in Adjunct Attachment”. In this paper they developed the Informative Boundary Hypothesis, a proposal according to which prosodic boundaries are interpreted not according to their individual size, which had been the leading mode of thought until

then, but rather with respect to each other, such that the phonetically-larger boundary would be processed as the prosodically and phonologically significant one, regardless of the size of the phonetically-smaller one.

To test their proposal, Clifton, Carlson and Frazier collected judgments on the intended meaning of a series of sentences with relative-clause attachment ambiguities, which had been recorded by a ToBI-trained linguist who alternated 0, ip and IP boundaries (as defined by ToBI conventions) at various break locations. The results showed a consistent shift in attachment decisions based on the difference between boundaries rather than depending on the absolute value of either, supporting their proposal. In 2002, Clifton, Carlson and Frazier further refined their hypothesis by testing a variety of syntactic structures traditionally considered to have attachment ambiguities², and replicated their findings with most structures.

In both papers, Clifton, Carlson and Frazier relied heavily on the definition of prosodic boundaries in ToBI, and recorded new sentences for each token, rather than manipulating a single base sentence, thus running the risk of introducing significant non-boundary phonetic detail into the sentences and therefore confounding the results. Furthermore, the large phonetic differences between the boundary levels included in the experiment, while useful in a preliminary test of the IBH, cannot offer insights into the finer-grained phonetic details of prosodic boundaries, and as such are not particularly useful in building a model of sentence processing and comprehension.

2 Specifically: Conjunctions with modifiers like “Old men and women with very large houses”; Possessives such as “Johnny and Sharon’s inlaws” and “The daughter of the Pharaoh’s son”; Relative Clauses like “I met the daughter of the colonel who was on the balcony”; Adverbials including “My uncle Abraham recited his poem naturally”; and Temporal PPs “Sammy learned that Bill called on Friday”.

0.1.2 Production Studies

Price et Al (1991) decided to investigate the cues to prosodic boundaries in the other direction, examining the behavior of boundaries in fully disambiguated sentences. They recorded FM radio announcers reading out loud five items each for seven different types of ambiguity³; the sentences were then extracted from the disambiguating context and annotated in a ToBI-like system consisting of seven separate levels (0 showing no break; 1 = prosodic word boundaries, 2 = “accentual phrase”, 3 = intermediate phrase, 4 = intonational phrase, 5 = groups of intonational phrases, breath intakes; 6 = sentence boundaries), and the phonetic properties around each boundary type were recorded.

Price et Al found a strong correlation between the lengthening of the phones preceding the boundary and the boundary size, a finding which was tested more rigorously in Wightman 1992⁴). They also found that pauses (intervals of silence), almost never occurred with boundaries of level 3 (intermediate phrase) or lower, suggesting that these are a strong cue to prosodic boundaries.

Based on the results of Streeter 1978, and Price et Al 1991⁵, the experiments in this thesis were designed to focus exclusively on timing-related cues, such as pre-boundary lengthening and pause duration, which have been found to be strong cues to

3 These were: Parentheticals, Appositions vs. Attached NPs, Main/main vs. Main/subordinate clauses; Tags; Far vs. Near attachment of Final phrase; Left vs. Right attachment of Middle phrase, and Particles vs. Prepositions.

4 Wightman et al conducted an in depth phonetic analysis of the Price corpora, and found that phone lengthening is confined to the rhyme (nucleus plus coda consonant) of the final syllable before the boundary. The increase in segmental duration appears to be linear, for the first four boundary stages (up to intermediate phrase boundaries), and then reaches a lengthening ceiling for boundaries of level 4, 5, and 6.

5 A number of studies were also carried out relating the phonetic cues of prosodic boundaries to other prosodic factors, such as rhythm and foot structure. Among these, Scott (1982) examined the role of duration as a cue to phrase boundaries, by manipulating the length ratio of two feet near the relevant phrase boundary. The results, now largely discredited due to a lack of interest in foot-based approaches to prosody, show that lengthened feet spanning the phrase boundary do often cue meanings associated with a phrase break at that location.

boundary location and also indicative of the existence of a boundary of a certain size. While intonation-related cues are doubtlessly important in real speech and are in fact crucial elements in models of prosodic boundary annotation such as ToBI, the assumption of this thesis is that pause and pre-boundary lengthening information by themselves would be enough to trigger the perception of a prosodic boundary and cue the appropriate meaning shift.

O'Malley et al (1973) specifically tested the cues present in disambiguating algebraic formulas, and found that pause duration was overall the strongest perceptual correlate (compared to pitch changes and vowel elongation) to boundary location⁶. Pauses (which they define as intervals of silence of 300 ms or above) correlate almost perfectly with perceived boundary locations, and the authors created a set of rules which used only pause information to reconstruct the formula structure, and in comparing it with human evaluations of the structure found that it was always at least 90% correct—despite not using information from pitch or vowel elongation.

Krivokapic (2006) also examined pause duration in the production of sentences with attachment ambiguities where constituents before and after the pause varied in length and complexity. She found that both constituent length and structure influence pause location, but while length effects are symmetrical (a large constituent both before or after the pause in question causes it to lengthen), only complex structures that occur after the boundary will affect the pause duration, actually shortening it.

In light of these studies, the use of pause duration as a key factor in the experiments in this thesis appears justified, although we will be including pre-boundary lengthening cues as well, to improve the naturalness of the synthetic speech.

⁶ This however was not the case when subjects produced the formulas at a fast rate of speech, when vowel elongation became the most strongly correlated cue, and silence the least.

0.1.3 Communication Experiments

Current work has largely moved on to studying the production/perception interface, and experiment designs have evolved to include two-subject game-like studies in which ambiguities, the knowledge of such ambiguities, and the knowledge of other's peoples differing world-view are manipulated to test which context-related factors affect the production of clear prosodic phrasing and its interpretation.

Snedeker and Trueswell (2002) tested whether speakers produced, and listeners used, disambiguating boundary cues in the absence of clear contextual ambiguity, and determined that while speakers only clearly disambiguate when they are aware of the contrast, listeners use the information whenever it is available.

Kraljic and Brennan (2005) tested the contexts under which speakers produce disambiguating boundaries, and whether this is influenced by the presence and/or disambiguation status of the addressee. They found that speakers do produce disambiguating cues regardless of whether ambiguity is present in the context, and regardless of whether they thought the situation was ambiguous for the listener or not. Furthermore, the listeners appear to always use the cues when they are available, even when the situation is contextually disambiguated.

The occasionally contradictory findings of these experiments suggest that much may depend on the specific properties of the experimental paradigm. One apparent constant, however, is that listeners do process prosodic information when it is available, and in fact even when it is redundant with the contextual information, so we are confident that subjects in these experiments will factor the prosodic information into the decision making process, even without being specifically instructed to do so.

0.2 Our Hypothesis

The model presented in this thesis takes Carlson, Clifton and Frazier's Informative Boundary Hypothesis (IBH) as a starting point, in assuming that the relative size of relevant prosodic boundaries is a factor in determining the interpretation of an ambiguous sentence. However, as we will be testing much finer-grained intervals (30 ms between tokens, instead of 300), we will be able to detect much smaller variability in the distribution of the results, and for this reason a slightly more refined model will be required, as will be described in the following sections. Testing multiple tokens for each item allows us to notice more subtle asymmetries in the distribution of results, as is described below.

0.2.1 The Experiment Setup

In this series of experiments, much like in Clifton, Carlson and Frazier (2002), we modify cues at two boundary locations that we have determined to be relevant for the desired ambiguity resolution. Each boundary location is modified to yield 7 distinct levels, and the intervals between them are 30 ms long, which incorporates both pause and pre-boundary lengthening information into a single scalar cue. As mentioned above, all tokens are generated from a single synthetic source and manipulated automatically.

Participants in the experiment judge every token within the 7-by-7 matrix, in an attempt to limit the variability between item responses. They select an answer corresponding to the interpretation (bracketing structure) which they believe is most appropriate, as well as a judgment of their confidence in their answer. These two scores are combined to yield a weighted response score, which is averaged across subjects for each token in the experiment.

The goal of this line of research is to find a model which predicts which structure is chosen for each token, which can be represented as a ordered pair (F, S) of the First and Second boundary levels—much like coordinates on a grid. This model assumes that the distribution of responses across the sets of ordered pairs (F,S) will be affected by the prosodic properties of the (F,S) pair, such that more extreme differences between the two boundary levels will yield more confident choices of either interpretation.

0.2.2 Overall Bias

Although we attempted to construct the most neutral sentences possible for the experiment, in many cases the two possible interpretations of the ambiguity were not equally likely to start with, resulting in bias in the data distribution, which we will refer to as Overall Bias.

Overall Bias is a type of bias, triggered by lexical, contextual or phonetic information, that changes the probability of the interpretation of each token within a matrix by the same amount.

Overall Bias can be triggered by any number of external variables, such as contextual information, including world-knowledge about the probability of various scenarios, or lexical choice and frequency, or even the phonetic properties of the sentence. For example, compare the most likely attachment point of the phrase *with the binoculars* in examples (1a) and (1b) below.

1. a. The birdwatcher saw the eagle with the binoculars.
b. The detective murdered the spy with the binoculars.
c. The detective saw the spy with the binoculars.

In 1a, the *with*-phrase has a high probability of being interpreted as an instrumental (late break interpretation), since birdwatchers are known to commonly carry binoculars which are used for seeing, and eagles instead rarely do. In the second case, however, binoculars are an unlikely murder weapon, and as such it may be easier to interpret the *with*-phrase as a modifier of the spy-NP (early break interpretation). In both cases, it would require very strong boundary cues in the other location to trigger the unlikely interpretation option, much stronger than for sentence (1c), where *binoculars* could just as easily modify the action of seeing as the NP *the spy*.

Sometimes the effect can be even more subtle, as in the case of the phonetic bias shown in example (2).

2. I want you to tap_the toy_with the feather

An equal amount of silence inserted at the two pause locations, indicated by an underscore, could be interpreted in radically different ways: the first location is between two stops, and as such part of the pause could be interpreted as either the closure of an unreleased preceding /p/, or as part of the initial closure of the following /t/, resulting in only a fraction of the silence being interpreted as true silence, and hence a boundary cue. In the second location, this kind of misunderstanding would not be possible, as both the pre-pausal /j/ and post-pausal /w/ glides have clearly discernible endpoints, and as such any intervening pause would have to be interpreted exclusively as silence, and hence as a boundary cue.

Overall Bias is so named because it affects all (F,S) tokens within the matrix equally, changing the probability of the interpretation of each (F,S) pair by the same

amount. In some cases, this can switch a token from receiving one meaning interpretation to the other, as demonstrated by the yellow dot in Figure 1⁷ below.

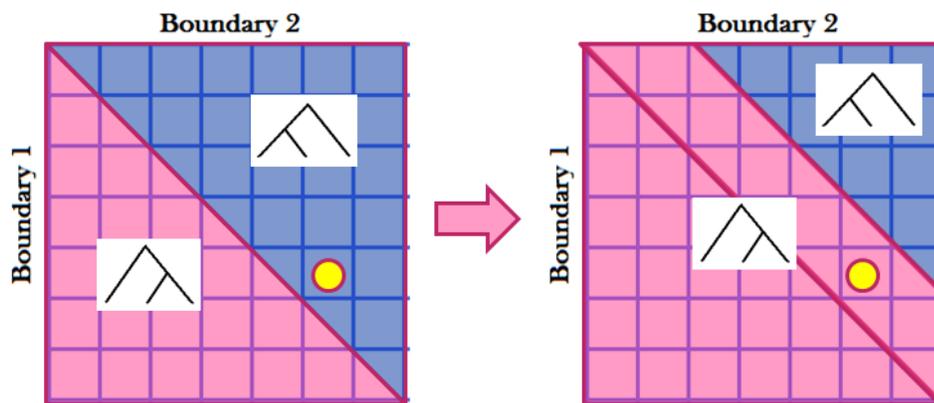


Figure 1: The Effect of Overall Bias on the Distribution Plot

It is also possible to plot the difference between boundary levels as a separate variable (named, aptly, Difference), with the two interpretations receiving positive or negative scores on the y-axis. The plots corresponding to the values in Figure 1, are shown in Figure 2: each bar in Figure 2 corresponds to a diagonal slice of the matrices in Figure 1, with the smaller bars within the clusters in Figures 2 (here represented as having different colors but the same height) represent the different tokens (boxes) within each diagonal.

⁷ In these Figures, the origin is at the top left corner, with the boundary sizes increasing as one proceeds downwards and to the right. The top-left corner box is thus point (0,0), top-right is (0,6), bottom left is (6,0), and bottom right is (6,6).

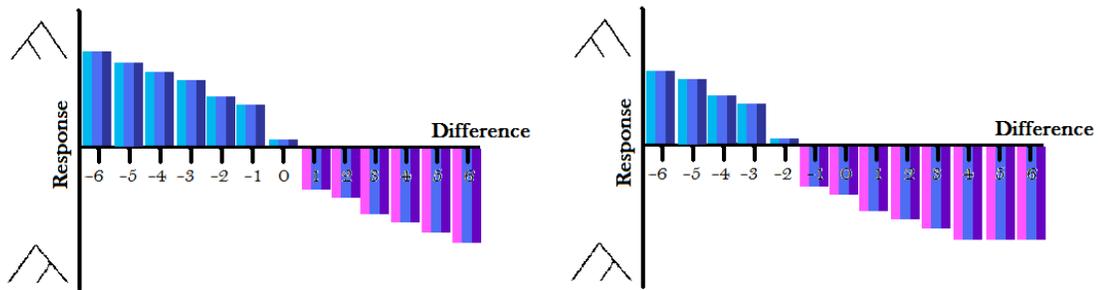


Figure 2: The Effect of Overall Bias on the Difference Variable

Note that a change in the level of the Overall Bias, where the contour line—the line between the two interpretations—remains parallel to the original line is translated onto the Difference Bar-chart as a shift of the intercept a few notches in either direction, but does not affect the distribution of the smaller bars within the clusters.

One crude way we propose to use to measure the presence of Overall Bias (which could be of either lexical, contextual, or phonetic origin) in the source file is to examine the average weighted response score given by subjects to the least-modified token⁸, the (0,0) soundfile. A strong skewness in the average response for this item would indicate that the item, despite being thought of as reasonably ambiguous by the experimenter, was really not perceived as such by experiment participants.

⁸ The (0,0) token is not exactly equivalent to the base file—the one generated by the synthesizer, manipulated by the experimenter, and then used to generate all the sound files in the matrix—since the (0,0) token has undergone manipulation that would leave the pre-boundary lengthening unchanged, but that would cut out the entirety of the pause silence. Testing other items in which there is pause silence would have them include potentially unnatural sounding pre-boundary lengthening, and of course the higher the manipulation number (in some cases levels 1 or 2 are enough), the more likely the presence of Conditional Bias would be, which would distort the Overall Bias effect we are trying to capture.

0.2.3 Conditional Bias

In our model of sentence processing, we also assume that listeners are at any point always trying to determine the meaning of the sentence, and to guess the structure of upcoming material. Under select conditions, this guesswork process can have strong effects the distribution of the responses, favoring one interpretation over the other, and we have decided to call it *Conditional Bias*.

Conditional Bias is created by giving Time at a *Point of Disambiguation*. A Point of Disambiguation is a point at which the preceding material can be assigned a meaning which has the effect that one of the two readings becomes relatively unlikely. Time (not only silence, but pre-boundary lengthening as well) is necessary for the material to be processed, and the processing of this material introduces information that makes alternative readings unlikely.

In an ambiguous sentence scenario like the one used in these studies, Conditional Bias reduces the probability of one or the other alternative interpretations being selected by varying amounts, depending on the properties of the individual tokens. However, it is important to note that it is not necessary for listeners to be aware of the ambiguity in order for Conditional Bias to take effect, and in fact this model could be applied to the processing of non-ambiguous structures as well.

In our experiment setup, we will be testing the presence of Conditional Bias at two boundary locations that flank the ambiguity. We predict that in cases of Conditional Bias, only one boundary—at the Point of Disambiguation—will be responsible for the choice of interpretations. In cases where both boundary locations

could be Points of Disambiguation, only the earlier one will have an effect, since we assume that listeners are trying to determine the sentence's structure as early as possible.

To give an example: assume that one hears the structure $3 + 4 * 5$ and has to calculate its meaning on the fly. A large pause after the first constituent, 3, would not really be very useful since there is nothing to be done at that point—this is not a viable Point of Disambiguation.

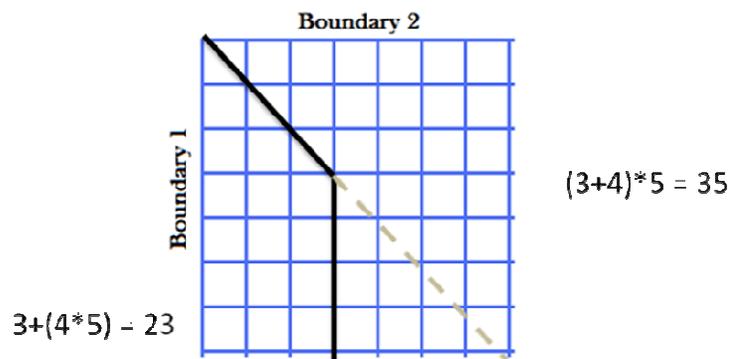


Figure 3: An Example of Conditional Bias in Action

A large pause after $3 + 4$, however, would allow for the calculation of the subtotal 7 at which point the bracketing would be $(3+4)*5$ for a total of 35. The distribution of responses is shown in Figure 3 below, and you can see that the Conditional Bias effect, here lowering the probability of an early break for certain tokens with a large Second boundary value, distorts the slope of the contour line. The predictions for each structure tested in this study will be discussed at the beginning of the relevant chapters.

Conditional Bias would therefore affect the distribution of responses within levels of the Difference variable (which are essentially diagonal slices of the grids

shown above). However, this effect would be systematic, showing a contrast of a particular level of the First or Second boundary factor within levels of the Difference boundary factor.

It is possible to confirm this graphically by separating the bars in the Difference variable bar chart into clusters, organized by First or Second boundary⁹ levels, and observing the behavior of the individual clusters. When the contour line of the Distribution plot cuts through the levels of the Difference variable, we expect to see systematic contrasts in the responses within and across clusters. Within the clusters affected by Conditional Bias, we expect a switch between interpretations as represented by the direction of the bars (up for late break interpretation, down for early break); and we expect this pattern to be systematically repeated throughout all clusters affected by the bias, as shown in Figure 4.

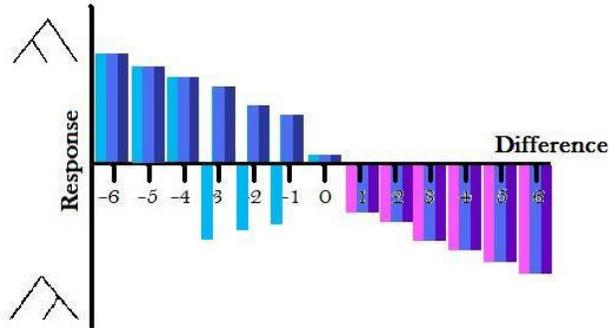


Figure 4: The Effect of Conditional Bias on the Difference Variable

⁹ Each difference factor level is comprised of a number of ordered pairs of First and Second boundary levels with the same difference between them: for example, Difference level 3 contains (6,3), (5,2), (4,1) and (3,0). Splitting this data across levels of the First boundary level will result in four groups (corresponding to First levels 6, 5, 4 and 3), and splitting the same data over levels of the Second boundary will produce exactly the same clusters, except labeled differently (3, 2, 1 and 0).

Of course, items can combine effects of the Overall Bias with the Conditional Bias, and all sorts of noise can distort the graphical representations of these response distributions, which is why we will be supplementing our findings with statistical analyses throughout.

0.2.4 Predictions of the Model

To summarize, our model predicts that the interpretation of sentences is determined by the relative size of the prosodic boundaries, but this can be affected by both Overall and Conditional Bias effects. Overall Bias simply shifts the point at which two boundaries are considered equal in favor of one or the other boundary, and is predicted to act uniformly across the tokens we are testing. Conditional Bias, on the other hand, only affects tokens for which one boundary is larger than a certain size, and makes the other boundary less- or un-informative for those tokens.

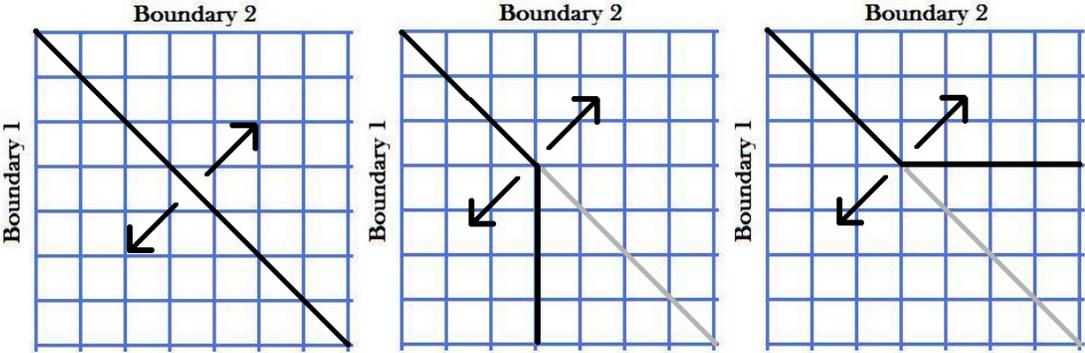


Figure 5: Predicted Contour Lines

The effects of Overall and Conditional Bias can be observed graphically in the shape and position of the contour line, which is the line that separates the different areas of interpretation on the Distribution Plot. Since Conditional Bias only affects boundaries larger than a certain size, we would expect the ‘plateau’ effects towards the bottom

and right of the plot only, while items with no Conditional Bias effect would show a simple diagonal line with an optional shift corresponding to the effects of the Overall Bias.

It is however possible for Overall Bias and Conditional Bias to Interact in ways that obfuscate their relationship. In Figure 6 we show a few scenarios combining Second boundary Conditional Bias with Overall Bias either favoring the early break interpretation (bottom-left portion expands), or late break (top-right portion expands).

A strong Overall Bias favoring an early break interpretation would be at odds with a Second boundary Conditional Bias (which favors the late break interpretation), and would result with a pervasive Second boundary Conditional Bias effect that slices through all levels of the First boundary, thus making the First boundary useless for the purposes of meaning determination, shown in Figures 6 a and b.

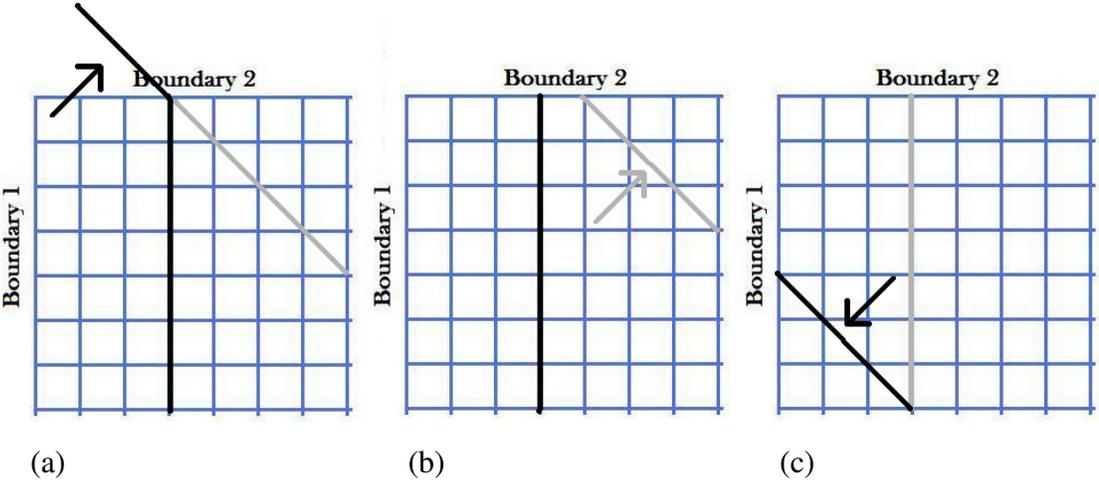


Figure 6: More Predicted Contour Lines

A strong Overall Bias favoring late break interpretation would instead result in complete redundancy, as the items for which the Conditional Bias would have triggered a late break interpretation (the right half of the distribution plot, with Second

boundary larger than a certain amount) are already assigned that interpretation by the Overall Bias, as in Figure 6c.

STUDY ONE: PHONETIC CUES

1.0 Introduction

In creating the stimuli for the experiments used in this thesis, we faced the difficult situation of having to approximate real speech as much as possible, while still using and manipulating cues that could be systematically controlled and statistically analyzed. The three cues isolated in previous literature as contributing to the interpretation of prosodic boundaries are Pitch, Pause duration, and Pre-boundary lengthening.

Because of our hypothesis testing the effects of Time on processing, as well as difficulties quantifying pitch, as well as issues of timing with respect to the syllables which carry it, we decided to focus on examining only silence insertion (henceforth *pause*) and pre-boundary lengthening (henceforth *duration*), which can both be precisely quantified (in ms duration) and manipulated (via Praat's PSOLA lengthening system).

At this point, however, it was necessary to take a short detour from the main narrative thread of this thesis, to confirm that both cues do in fact contribute to meaning determination, and that the presence of both cues is more informative than the use of either cue on its own.

1.1 Method

This study focuses on manipulating the structure “B | plus C | times D ”, the baseline item to be discussed more in depth in the next chapter, with respect to pause only, duration only, and pause+duration cues. The structure of this experiment is slightly simpler than the ones forming the core of the thesis.

1.1.1 Stimuli

In order to introduce as little variability as possible across tokens, all the sound file sets used in the experiments in this thesis were generated from a single base token via an automated Praat script. This base file was in turn synthetically generated using a demo version of Eloquent Technology's Elocutor formant-based synthesizer, and was hand-corrected to remove any remaining prosodic information that would create unnecessary bias in the base file. Pre-boundary syllables were compared to versions produced by the same synthesizer in non-boundary locations, and shortened if necessary. Pauses at the two relevant boundary locations were cut to a maximum of 25 milliseconds, and preceding consonant burst releases were also trimmed to fit with the shorter boundary, when their length caused them to sound unnaturally spliced, in an attempt to make the base item as neutral sounding as possible.

A textgrid was then created for the base file, which targeted 20 ms of the pause intervals, and 30 ms of the steady state of the pre-boundary syllable nucleus, for automated lengthening. The base file and textgrid were then fed through the Praat script that used Praat's PSOLA manipulation function to systematically lengthen the targeted section in fixed increments as specified by the user, and flatten the pitch for the utterance to 100 Hz.

For this study, the pause-only and duration-only cues were modified in four steps (from level 0 to 4) of 22.5 ms increments, for a total of 90 ms maximum lengthening at each boundary location. This resulted in a 5 by 5 matrix of possible manipulations, once each for duration-only and pause-only stimuli. The pause-and-duration tokens were instead modified by adding six increments of 30 ms each (from level 0 to 6), for a total of 180 ms maximum lengthening at each boundary location, which in turn was divided equally between the pause and duration cue. The most extreme manipulations (0,6), (6,0), and (6,6), and in one case (0,0) as well, were

thought to be uninformative at this stage, as the interest lay primarily in determining the extent and location of the cross-over between interpretations, and as such were removed from the experiment, resulting in a total of 45 tokens for the 7 by 7 manipulation with both cues, and 22 each for the pause-only and duration-only items¹⁰.

1.1.2 Subjects

Six subjects were recruited from the Cornell undergraduate population, whose only requirement was that Standard American English be their native language. Subjects were compensated for their time with \$5 or one extra-credit point in an undergraduate psychology class.

1.1.3 Experimental Setup

In this first study, subjects were asked to make a decision between three possible prosodic structures: $B+(C*D)$ (early break interpretation), $B+C*D$ (flat prosody), and $(B+C)*D$ (late break interpretation), with the formulas displayed in the same order as shown here. The sentences received a score of 1 for early break interpretation, 2 for flat structure prosody, and 3 for late break interpretation, which were used to calculate the statistics for each item and variable.

Participants were also familiarized with audio tokens of prototypical items for each of the three prosodic structures (with both pause and duration cues present) in a training session before the beginning of the study. Subjects heard the sound files through headphones in a sound proof booth but were instructed to abstract away from the unnaturalness of the synthetic speech as much as possible.

¹⁰ In later experiments this assumption was judged to be incorrect and the whole matrix of factor levels was tested, resulting in 49 sound files for the 7 by 7 standard format.

1.1.4 Variables

In the experiments in this thesis, a number of statistical analyses will be run to analyze the distribution of the dependent variable (Response) for the different items, and whether and how this is affected by three independent variables: First, Second, and Difference. First and Second are quite transparently simply the manipulations applied to the two boundary locations, which are reported for each item, and each are comprised of a number of levels (typically 7), representing discrete and consistent manipulations. Thus, a boundary level of 3 will have the same phonetic properties across all different sentences (except when otherwise noted) as well as across locations (First or Second position), and is the same phonetic distance from a level 2 boundary as from one of level 4.

The third independent variable, Difference, captures the degree of phonetic separation between the levels of factors First and Second. The variable is calculated by subtracting the level number for the second boundary from that of the first: two boundaries of equal size would therefore have a Difference value of 0; tokens with a larger first boundary than second would have a positive Difference value, and tokens with a larger second boundary than first would have a negative Difference value. The exact millisecond difference between the boundaries can once again be calculated by multiplying the Difference value by the interval between First/Second boundary levels, which is reported separately for each item.

The Response variable for this experiment is a three-level coding of subjects' choices: 1 for early break interpretation, 2 for flat structure, and 3 for late break interpretation. The individual subjects' ratings were averaged for each token, and the resulting score was used as a basis for the statistical analyses and comparisons run in this chapter.

It is important to remember that statistical significance does not imply causation, but simply a correlation between the levels of a factor and a variation in the distribution of the Response variable. Tests must also be run to ascertain whether the significance is consistent across all of the data, for example within subgroups as defined by levels of other factors; as well as whether the distribution of responses is constant (non-significant) within the levels of what is hypothesized as being the explanatory factor. Furthermore, the distribution must also correspond to the linguistic predictions, such that neighboring factor levels behave more similarly than distant ones.

1.2 Results

When both cues are applied to the same sentence, the division between prosodic interpretations is extremely clear, with both boundaries achieving significance at $p < 0.001$ (and $F > 14.0$), with no interaction between factors ($p = 0.296$).

Tokens modified by only the Duration cue are similarly highly significant ($p < 0.005$, $F > 4.0$), with no interaction between factors ($p = 0.386$). However, the tokens modified using only pause insertion were significant only with respect to manipulations of the First boundary ($p = 0.013$, $F = 3.337$), and only approach significance for the Second boundary factor ($p = 0.089$, $F = 2.073$). There is no interaction of factors here either ($p = 0.287$).

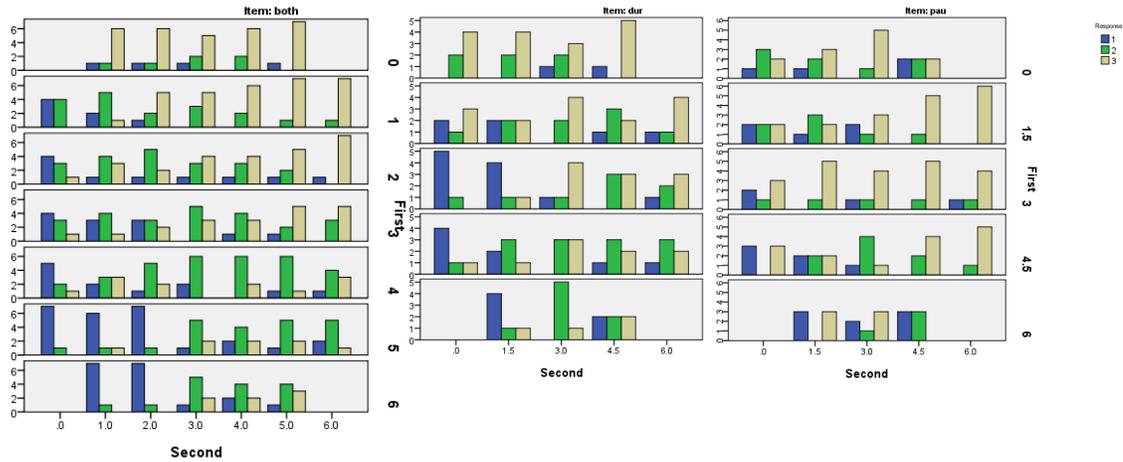
The Difference variable is highly significant for tokens manipulated with both P+D cues as well as those manipulated only with pre-boundary lengthening ($p < 0.001$), and approach significance when manipulated only with pause cues, at $p = 0.054$, with $F = 2.128$.

1.3 Discussion

The main point of this study was to compare cues across items, rather than to start building a model for the processing of prosodic boundaries, the results we are interested are the magnitude of the effect of the same factors across different cues, such that the strongest cue or combination of cues would be selected for manipulations in the rest of the thesis.

The results show that duration (pre-boundary lengthening) is a better cue than pause, when they are used alone. The pause tokens appear to have an overall late boundary bias, and as can be seen in (a) Duration+Pause (b) Duration-Only (c) Pause-Only

Figure 7, it is difficult to find tokens for which all speakers agree on the early boundary interpretation. The duration tokens are more evenly distributed across prosodic forms, and listeners find the tokens to provide clear examples of all three available prosodies.



(a) Duration+Pause (b) Duration-Only (c) Pause-Only

Figure 7: Response Distribution for Study One Items

While tokens which are manipulated exclusively with respect to duration (pre-boundary lengthening) do achieve significance at both boundaries in this context, the addition of pause cues does reinforce the strength of the effect, as can be seen in (a) Duration + Pause (b) Duration-Only

Figure 8 from the larger range of values and the plateaux reached for both early and late boundary values, which are not present in the duration-only modified response values.

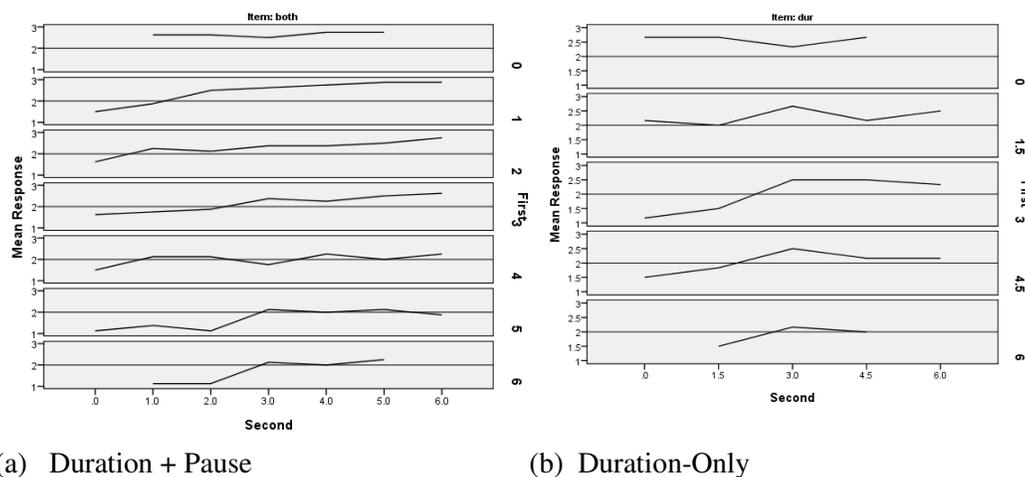


Figure 8: Comparing Maximum Mean Response Displacements

For this reason, we decided to use both cues in the manipulation of the tokens for the remainder of the experiments. It is unclear whether the weakness of the pause results (where the second boundary not significant) can be generalized to natural speech as well, or whether it is a reflection of unnaturalness of synthetic speech, where pauses may be more often interpreted as glitches or hesitations, and lengthening is seen as a more reliable cue to boundary position.

STUDY TWO: SIMPLE CONJUNCTIONS

2.0 Introduction

In this and following sections, we aim to calculate the effect of variations in prosodic boundaries on the interpretation of the sentence in question. This first set of experimental items was designed to be as simple, balanced, and semantically and structurally neutral as possible, so as to provide a good baseline for comparison across items.

The results show that both boundary factors as well as the Difference variable are significant for most items, but the variability of the response within each level of the factors suggests that the explanation is more complex than just simple phonetic differences between boundaries.

Our Conditional Bias predictions proved correct for some items, while for others our processing model had to be modified in order to account for shortcuts introduced by the experimental setup.

2.1 Method

In this section we will describe the method by which the experiment was run, including the stimuli creation and manipulation, the role and number of the participants, the setup of the experiment and the details of the experimental task, as well as the Conditional Bias results predicted by our processing model (we always assume that the items chosen display little or no Overall Bias, and as such have no predictions for that element).

2.1.1 Stimuli

The first items tested were constructed to be as structurally neutral as possible, in order to form the baseline against which other items would be compared. The

stimuli were designed with a symmetric tripartite structure connected by simple conjunctions, and each of the three constituents was designed to be both phonetically light (monosyllabic, with simple internal structure and comparable length in pronunciation), as well as semantically uninteresting.

Four items were tested in this study:

3. a. B | plus C | times D
- b. B | and C | and D
- c. Rose | and Steve | and Kim
- d. Eve | or Jude | and Sue

The base sound files, generated by the Eloquent synthesizer, were modified via an automated script to have both pause insertion and pre-pausal lengthening at the target boundary locations. A set of 49 sentences was constructed for each item, varying the first and second boundaries in 7 steps and crossing both factors. Each interval between factor levels consisted of an increase of 30 ms (from 0 to 180 ms total), which was equally divided between silence (called *pause*) and pre-boundary lengthening (called *duration*).

To verify the accuracy of some early results, both conjunction items 3b and 3c were retested with larger duration increments between factor levels. The new items had a maximum of 360 total lengthening (cf. 180 for the previous items), divided over 6 steps of 15 ms lengthening and 45 ms pause, resulting in a 1:3 ratio between duration and pause amounts (maximum 90:270 ms lengthening), rather than 1:1 as was employed for shorter manipulations. The reason for this is that lengthening a vowel by more than 90 ms (in this case, up to 180ms) resulted in strongly unnatural-sounding tokens, and given the use of synthetic, pitch-flattened tokens, it seemed unnecessary and undesirable to add more unnaturalness to the sound files. Long pause durations (up to 270 ms) on the contrary seemed to be better accepted.

The resulting sound files were then randomized and interleaved with fillers so as to avoid prompting attachment bias; subjects were tested on all items of a set (49 sound files) to minimize inter-item inconsistencies.

2.1.2 Subjects

Participants were recruited from the Cornell undergraduate community and participated for \$5 or one extra credit point for their psychology class. The items in this thesis were run at different times and in different blocks, and as such were often tested on different numbers of subjects, although we decided following the first results that 6-10 subjects would be enough to guarantee that a strong effect would be picked up by the statistics. Results could be qualitatively observed (without the statistics) with as few as four subjects.

Twenty subjects were run on item (1a); six on item (1b) in its short form and five on the long form; four on item (1c) in the short form, and seven on the long form; and seven more on item (1d).

2.1.3 Experiment Setup

Several changes were applied to the experimental paradigm that had been used in Study One to make it more sensitive to the hypothesis being tested. Firstly, subjects were presented with a forced choice between two possible answers, early or late boundaries, representing the different available syntactic structures and of the ambiguous sentence (instead of three choices corresponding to three prosodic structures). These judgments received a score of -1 for early boundary and 1 for late boundary position, and were averaged across subjects for each token to yield the Simple Response score.

Subjects were also asked to rate their confidence in their decision on a scale from 1 (unsure) to 5 (certain), which was converted to a goodness rating from 0 to 4, and multiplied by the attachment score to yield the Weighted Response score (referred to as Wresponse from now on). Items with low confidence scores would therefore get a score of 0, wiping them out of the statistical calculations, while items with larger confidence ratings would get a score of 4 and -4, thus affecting the mean and distribution of the Wresponse score to a larger degree.

Lastly, the audio training session was replaced by a written set of instructions in which the items were disambiguated in contexts, and participants were encouraged to ask the experimenter for clarifications, but they were not exposed to any audio rendition of either interpretation to avoid creating any expectations or bias. The experiment sound files were still presented to participants through headphones in a sound proof booth, and subjects were instructed to abstract away from the unnaturalness of the synthetic speech as much as possible and focus only on the intended interpretation.

The specific answer choices that were presented to participants varied slightly from item to item. The formula (3a) was presented visually with both possible bracketing structures explicitly shown, and subjects were simply asked to pick which one they felt was intended.

4. a. $(B + C) * D$
- b. $B + (C * D)$

Item (3b) required a short setup in order to gain plausibility: in the written training materials subjects were told that the police was investigating two crimes that had taken place the previous night, and the investigations had narrowed the search to three suspects, B, C and D, of which two had worked in a pair, and one alone. The

answer choices consisted of the same conjunction, with “together” and “alone” following the two major constituents to emphasize the interpretation already conveyed through the use of punctuation.

5. a. B and C together; and D alone.
- b. B alone; and C and D together.

Similarly, for the third item (3c), subjects were told that these three people were arriving at a party, but two of them were a couple, and the third just rode in the same car with them. The answer choices were disambiguated on screen as follows:

6. a. Rose and Steve, a couple; and Kim
- b. Steve and Kim, a couple; and Rose

Item (3d) had a similar setup, where subjects were asked who the computer voice was speaking about. In order to help with the disambiguation, the answer choices were prefaced by *either* or *both*, to force the relevant conjunction to have wider scope:

7. a. Either Eve; or Jude and Sue
- b. Both Eve or Jude; and Sue

2.1.4 Conditional Bias Predictions

The items tested in this section are predicted to have Second-boundary Conditional Bias. Much like the numerical expression $3 + 4 * 5$ discussed in Section 0.2, there is nothing to process at the First boundary location (after *B* or a single name), and so a large boundary doesn't help with wrapping up or processing the structure up to that point. A large boundary at the second location (after *B plus C* or *Eve or Jude*, for example) would however allow for the time required for that structure to be built: in the numerical example, this would allow the calculation of the intermediate total. The remaining part of the token would then most probably be added on top of that, in a high-attachment or late break structure. Thus, a long boundary at

the Second boundary location (Second Boundary Conditional Bias) would decrease the probability of an early break interpretation.

2.2 Results

For each item, we report on a number of statistical tests run to assess the influence of the First, Second, and Difference factors on the distribution of the Wresponse (weighted response) score. These include the standard Analysis of Variance (ANOVA), as well as the Multivariate Analysis of Variance (MANOVA), which performs a series of ANOVAs for one factor within levels of the other. The factors analyzed here (First, Second, Difference) are considered Fixed for statistical purposes unless otherwise stated; and significance was set at $\alpha = 0.05$ throughout.

However, it must be kept in mind that due to the limited number of subjects run on some of the items, statistical tests may not always catch all or any of the patterns in the data. For this reason we supplement the statistics with a series of figures displaying the distribution of Wresponse scores for First, Second, and Difference boundary levels, in which we can observe finer patterns and tendencies.

2.2.1 Algebraic Formulas

For the Formula conjunction (3a), both factors First and Second were highly significant: $F = 17.314$ and 56.785 respectively, and $p < 0.001$ in both cases. The interaction of First and Second was not significant ($F = 0.982$ and $p = 0.500$). MANOVA tests of the First and Second boundary factor levels show that the Second boundary is significant within every level of the first ($p < 0.005$ throughout), but the First boundary is significant only within the first four levels of the Second boundary (levels 0-3), at $p > 0.014$; when the Second boundary level is 4, 5, or 6, the First boundary has no effect on the weighted response (Wresponse) distribution. These

results can be visualized in Figure 9, where the contour line separating the early and late break attachment decision areas clearly shows a skewedness at high levels of the Second boundary factor.

The variable Difference is significant with $F = 32.960$ and $p < 0.001$, but a MANOVA test of the effect of First/Second boundary levels¹¹ within the levels of the Difference boundary showed significance at $p < 0.02$ (and often $p < 0.001$) for all items except the Difference level -1 (corresponding to when the First boundary is one step smaller than the Second, such as the pair (4,5)), where $p = 0.197$.

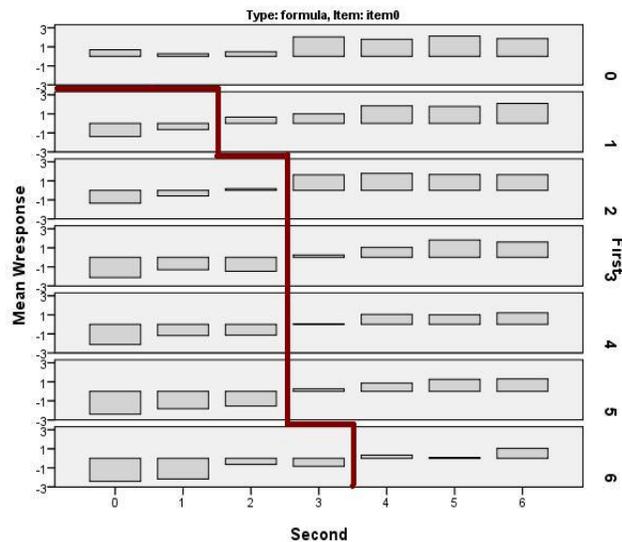


Figure 9: Distribution Plot for item "B plus C times D"

¹¹ Each difference factor level is comprised of a number of ordered pairs of First and Second boundary levels with the same difference between them: for example, Difference level 3 contains (6,3), (5,2), (4,1) and (3,0). A MANOVA test on the effects of the First boundary within this level of the Difference variable will divide the data into four groups corresponding to those with First boundary level 6, 5, 4 and 3; but a MANOVA test for the effects of the Second boundary will also split the data into the same four groups, corresponding to Second boundary factor levels 3, 2, 1, and 0. Although labeled differently, the groups, and thus the analyses, will be identical regardless of whether the First or Second variable is used, and for this reason it is impossible to extract the effects of either, and I will be referring therefore to the joint effect of the "First/Second" factors.

One other way of quantifying the effects of Overall Bias is to look at the score of the token at (0,0): if this token receives a Mean Wresponse score that strongly deviates from zero—which we have arbitrarily defined as a score smaller than -1 or larger than 1, out of a maximum of 4 and minimum -4 —this could be indicative of a strong Overall Bias present in the base file. However, it is important to keep in mind that since we are relying on the score of a single token, which in turn is decided by as many votes as subjects that were run on the item, this datum can be easily affected by outliers in individual subjects’ response assignments¹². For this item, the Mean Wresponse score of token (0,0) was 0.69, which under our assumptions does not indicate strong bias in the base file.

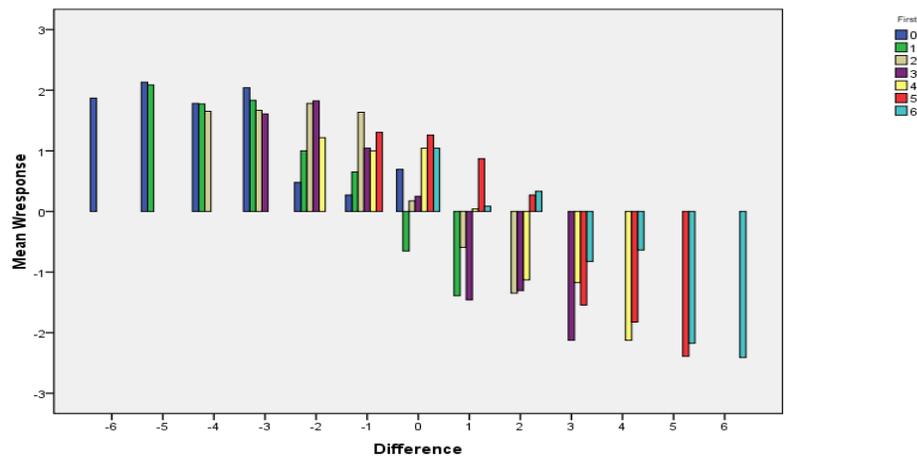


Figure 10: Difference Bar-chart for item "B plus C times D"

¹² We also considered analyzing the score averaged over all of the items with Difference = 0, that is (0,0), (1,1), (2,2), etc., as this should—in a case of Overall Bias—provide a stronger measure of the bias affecting items with the same phonetic prosodic boundary sizes at both locations, and would be less easily affected by outliers. However, examining larger boundary sizes exposes us to the risk of interference from Conditional Bias effects, and it was decided that it would be better to stick to this flimsier, but more accurate, representation of the bias in the base file.

These results are echoed in Figure 10, where we can observe strong variation in scores within the clusters representing each level of the Difference factor. Although what may appear most striking is the difference in height between bars in the same cluster, what is key to our comparison of Overall vs. Conditional Bias is the presence of systematic alternations from the positive to the negative planes of the y-axis, or vice versa. In this figure, note the contrast between the rightmost two bars and the others in Difference levels 0, 1 and 2.

2.2.2 Suspects B and C and D

The simple conjunction item (1b) also showed significance for both boundaries: $F = 9.754$ and 2.966 , and $p < 0.001$ and $p = 0.008$ for First and Second respectively. The interaction of First and Second was not significant ($F = 0.891$ and $p = 0.651$). MANOVA tests of the effects of the First boundary within the levels of the Second showed significance when the Second boundary was 1, 3, 4, and 6; the Second boundary was similarly significant within the levels of the First when First was 0, 1, 2, 3, and 5.

Figure 11 illustrates the picture more clearly: while both Boundaries have a strong effect on the distribution of the Wresponse score, even within most levels of the other factor, the contour line shows that the point at which the interpretation switches from early to late break is more strongly affected by the First boundary factor levels (and this is reflected by the relative strength of the statistical significances, which can be obtained by comparing F values).

The mean Wresponse score for token (0,0) was 0.714 for this item, which does not suggest the presence of strong Overall Bias effects in the base file. Although the contour line, as drawn in Figure 11, suggests the presence of a slight Overall Bias towards a late-break interpretation, the low Confidence values (thin bars) for tokens

(0,0) and (1,0) are such that the line could fall anywhere around those points, and what is drawn here represents the strictest reading, which can however be influenced by noise in the results.

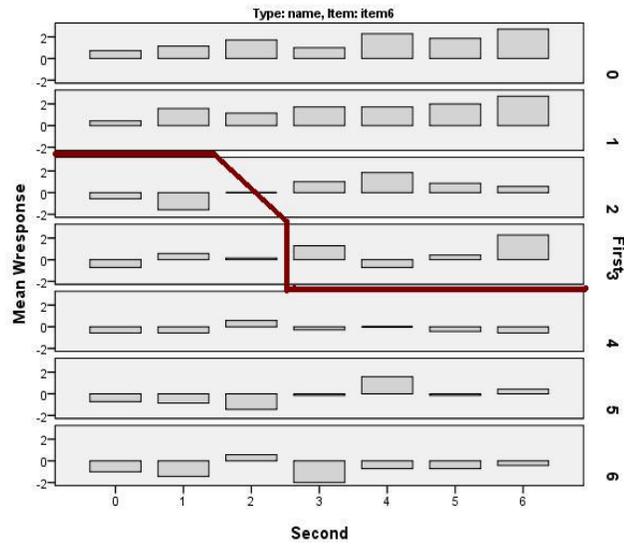


Figure 11: Distribution Plot for item "B and C and D" (short)

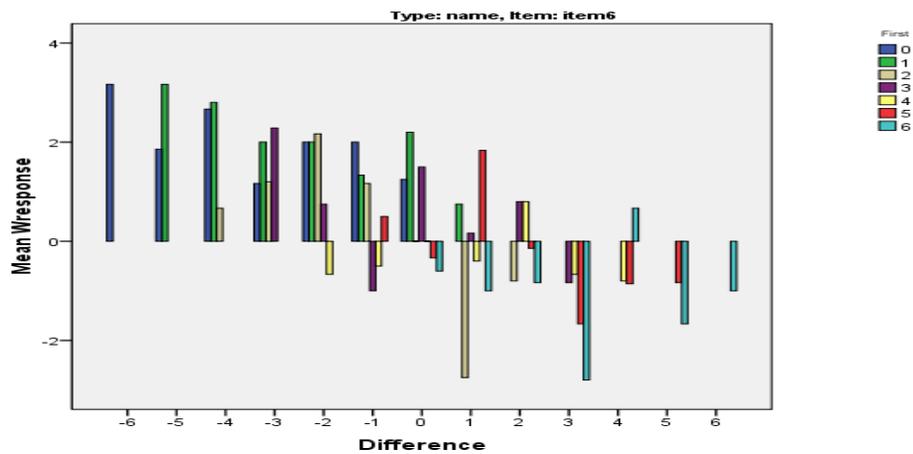


Figure 12: Difference Bar-chart for item "B and C and D" (short)

The Difference variable was strongly significant for this item ($p < 0.001$, $F = 35.473$). MANOVA tests of the effects of First and Second boundary variation within

levels of Difference were significant at $p > 0.05$ for when Difference was -6, -5, -2, and 5, but the variability captured by this analysis does not appear to be systematic, as shown in Figure 12. Note instead the systematic contrast between the first three bars and the others in Difference levels -2, -1 and 0, corresponding to the perfectly horizontal contour line at between levels 3 and 4 of the First variable in the previous figure.

The same item was also run with longer phonetic increments per step (6 steps of 60 ms each, divided in a 3:1 ratio between pause and pre-boundary lengthening), and these showed strong significance for both items ($F = 19.945$ and $p < 0.001$ for the First boundary, and $F = 17.611$ and $p < 0.001$ for the Second), with no interaction of factors ($F = 0.921$ and $p = 0.601$). The MANOVA tests of the effects of the First boundary within Second showed strong significance ($p < 0.05$) for levels 2, 3, 4 and 5 of the Second boundary, as well as for Second with First at levels 1, 2, and 3 ($p < 0.01$).

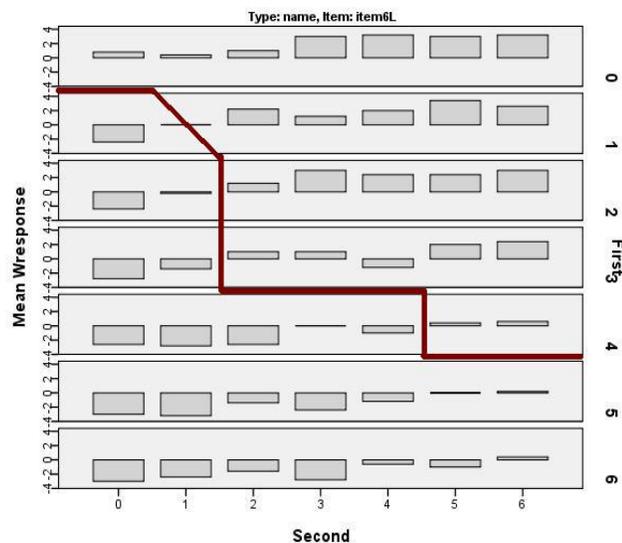


Figure 13: Distribution Plot for item "B and C and D" (long)

The Mean Wresponse score for token (0,0) was 0.8 here (compare to 0.714 for the same token when it was part of the shorter phonetic manipulations), which does not indicate the presence of strong Overall Bias in the base file.

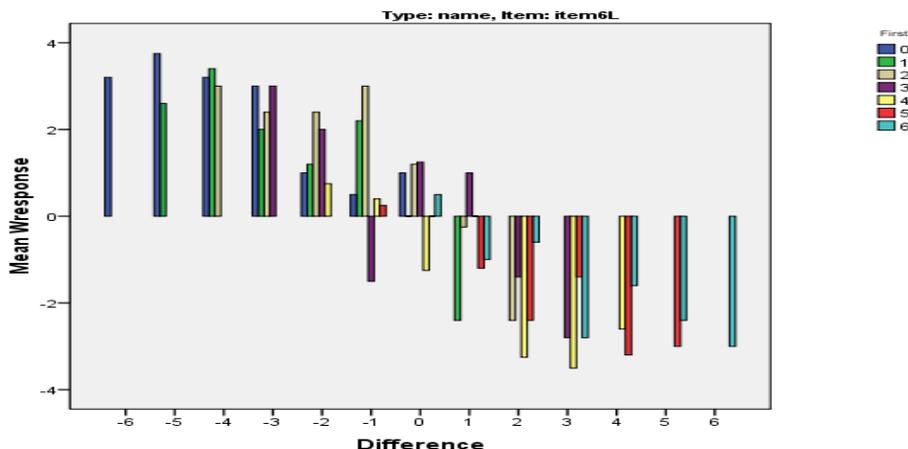


Figure 14: Difference Bar-chart for item "B and C and D" (long)

The variable Difference was also highly significant ($p < 0.001$, $F = 18.356$), and MANOVA tests of the effects of First/Second boundaries within levels of Difference were significant at an alpha of 0.05 for all but levels -2 and 0: however, no systematic difference can be detected on this chart (consistently opposing scores at opposite ends of the clusters), as was visible in the previous two Difference figures, suggesting that Conditional Bias may only have a limited scope in this item.

2.2.3 *Rose and Steve and Kim*

Item (3c), a conjunction with monosyllabic names, displayed significance only for the First boundary manipulations ($F = 5.668$, $p < 0.001$), but not for the Second ($F = 1.589$, $p = 0.149$); or for interaction within the factors ($F = 0.573$, $p = .978$). Analyzing the data further under a MANOVA analysis shows that the First boundary

was significant ($p = 0.033$, $F = 2.32$) only within level 4 of the Second boundary. The distribution of Mean Wresponse scores is shown in Figure 15, which shows that while there is a sizeable area that displays early break interpretation, these rarely are confident scores (note the thinness of the bars).

The mean Wresponse score for token (0,0) is 0.875, which according to our assumptions should not indicate the presence of a strong Overall Bias effect in the base file.

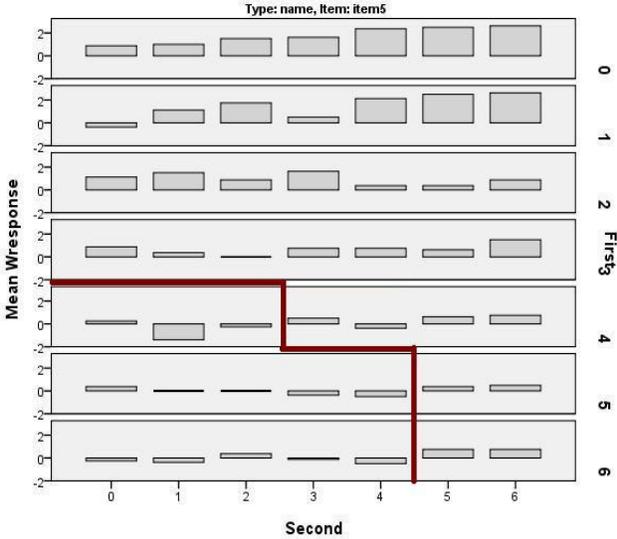


Figure 15: Distribution Plot for item "Rose and Steve and Kim" (short)

The Difference variable is highly significant for this item ($p < 0.001$), and a MANOVA analysis shows that First/Second have significant effects on the distribution of the Wresponse variable only within levels -6, -5, and -4 of the Difference variable: a strong contrast is clearly present within the cluster for Difference level 2, but this cannot be extended to adjacent clusters, suggesting that it might not correspond to a true effect of Conditional Bias.

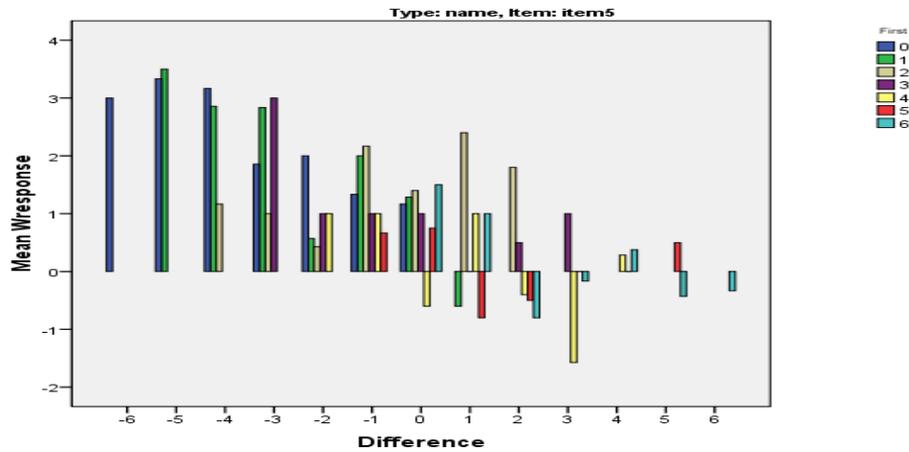


Figure 16: Difference Bar-chart for item "Rose and Steve and Kim" (short)

This item was also tested with longer intervals (six 60 ms steps divided in a 3:1 ratio between pause and pre-boundary lengthening) to gain further insight into the results collected thus far. With longer intervals, the significance increased to $p < 0.001$ for both the First and Second boundary factors, and MANOVA analyses showed significance at $p < 0.006$ for the First boundary within all levels of Second, and with an alpha of 0.05, the factor Second was significant within levels 1, 2, 3 and 4 of the First boundary. Figure 17 clearly displays this asymmetry, which reflects First-boundary Conditional Bias.

The Mean Wresponse score for the token (0,0) is approximately 0 for this item—a good result, in that it clearly shows that there is no bias present in the base file, but surprising, given that the identical token in the previous testing round received a score of 0.875. However, as both fall under our threshold for Overall Bias (set at 1, out of a possible maximum score of 4), there is no cause for concern in the variation between items. As mentioned previously, the exact value of this particular token, which is calculated by averaging the Wresponse scores across all subjects, is much

more strongly responsive to the effects of outliers when the sample size is small, as in the case of the lengthened version of this and the previous item (3b).

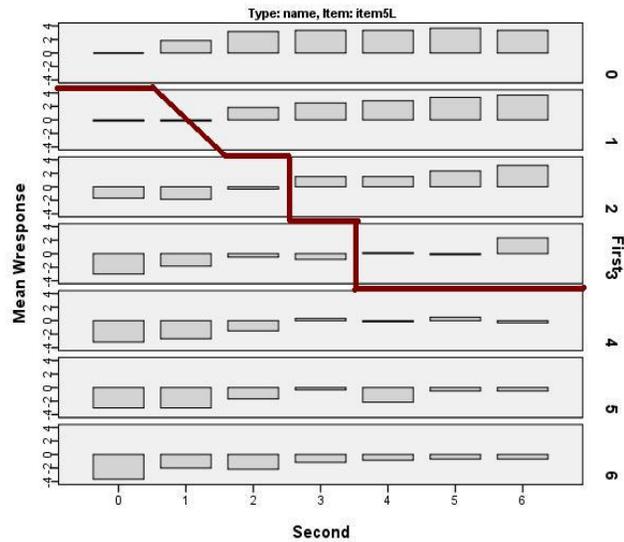


Figure 17: Distribution Plot for item "Rose and Steve and Kim" (long)

The variable Difference is highly significant ($p < 0.001$, $F = 35.500$), but a MANOVA analysis shows that there is significant variability ($p > 0.001$) within the levels of Difference which can be attributed to the effect of First/Second boundaries, with the only exceptions of the Difference-levels 0 and 2, which are not statistically significant. However, as was noted before, this statistical significance tends to capture random noise more accurately than systematic contrasts, which, although very slight, can be noted within Difference levels -2 and -1.

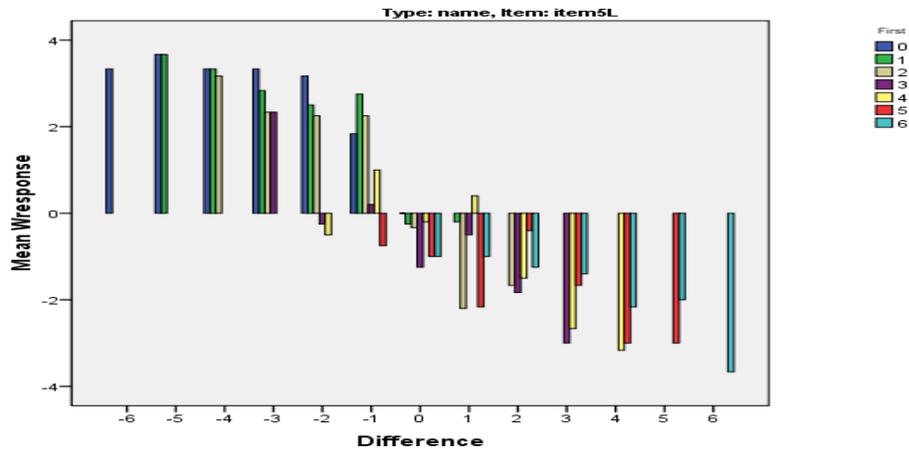


Figure 18: Difference Bar-chart for item "Rose and Steve and Kim" (long)

2.2.4 *Eve or Jude and Sue*

Item (3d) failed to show significance for either the First or Second boundary factors ($p = 0.127$ and $p = 0.434$ respectively), with respect to the Weighted Response variable. When considering only the distribution of the simple response variable, which tallies subjects' scores but not how "good" they judged the response to be, the First boundary approaches significance at $p = 0.066$ ($F = 2.002$), but the second boundary remains strongly not significant ($p = 0.386$, $F = 1.062$). No interaction of factors was reported ($p = 0.984$, $F = 0.548$). Given the non-significance of the First and Second boundary factors overall, it is not surprising that no boundary resulted significant within any level of the other, with $p > 0.300$.

The Mean Wresponse score for token (0,0) of this item was 0.66, which again does not reach our threshold for representing Overall Bias effects in the base file.

Figure 19 below shows the distribution of results and the contour line separating the zones of different interpretations, and as with item (3c), the lack of statistical significance is probably due to the low level of confidence (and therefore smaller Weighted Response score) of these early break items.

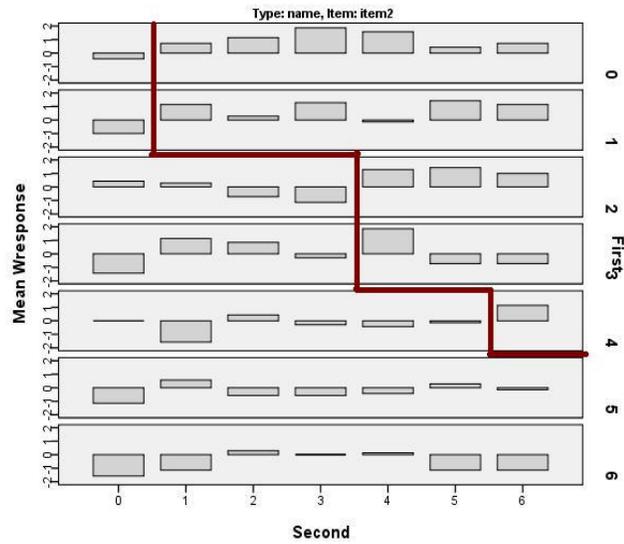


Figure 19: Distribution Plot for item "Eve or Jude and Sue"

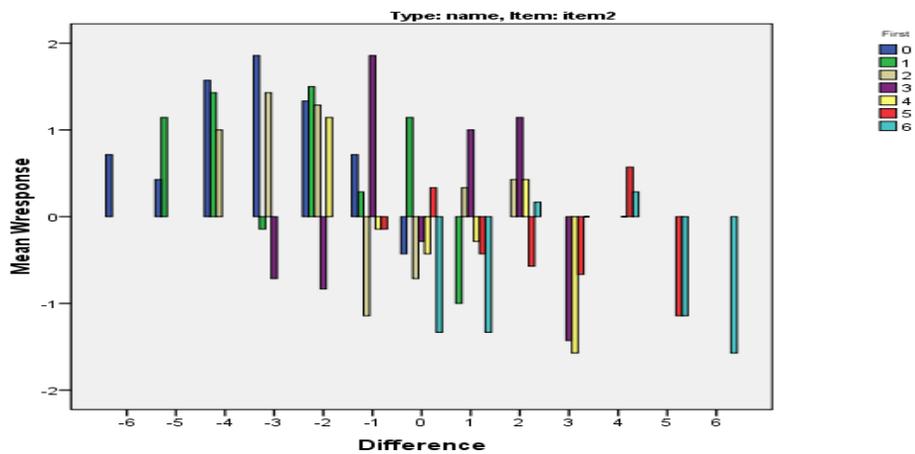


Figure 20: Difference Bar-chart for item "Eve or Jude and Sue"

The variable Difference approached significance using both the weighted (Wresponse) and unweighted (Response) dependent variables, at $p = 0.08$ and 0.054 respectively. There was no significance in the MANOVA analysis of First/Second within any level of the Difference variable (which is not surprising given the overall

lack of significance), but the distribution of Mean Wresponse by Difference in Figure 20 appears to show only random fluctuations.

2.3 Discussion

For all items in this section, the distribution of the Wresponse variable can be shown to depend on the size of both the First and Second boundaries. However, a closer look at the Difference variable's effect on Wresponse, through MANOVA analyses examining the effects of the First/Second factor within levels of the Difference factor, clearly shows that raw phonetic difference between boundary sizes is not the sole factor responsible for the subjects' choice of interpretation.

As predicted by our model of Conditional Bias, the item *B plus C times D* showed a clear Second boundary Conditional Bias, as predicted by our processing model, which could be noticed both in the contour chart as well as in the Difference distribution bar graph.

The item *Eve or Jude and Sue*, although predicted to have a Second boundary Conditional Bias, does not display it in this data set. The Distribution Plot, as well as the Difference Bar-chart, do not show any systematic preference of one boundary over the other. The contour graph shows a very slight Overall Bias—not enough to even show up in our crude assessment of the base file bias based on the (0,0) token results—which appears to slightly favor the early break interpretation overall, meaning that two equal boundaries are perceived as having a First boundary that is more salient than the Second. Conversely, in order for two boundaries to be perceived as equal (at the Contour Line), the Second boundary would have to be slightly larger than the First. This is quite possibly due to the fact that the First boundary is located after a voiced fricative (“Eve”), whereas the Second is located after a voiced stop (“Jude”),

and as such part of the Second boundary may have been interpreted by subjects as being part of an unreleased stop closure, for example.

Contrary to our predictions, however, both items *B and C and D* and *Rose and Steve and Kim* display a clear First-boundary sensitivity threshold, after which the probability of a late break interpretation is reduced, even in situations in which the Second boundary is phonetically larger than the First. Although the items in this section should have identical syntactic structures, we suggest that the difference in Conditional Bias results is due to the different nature of the tasks that participants are asked to complete, as will become clearer in later chapters.

The answer choices for these items explicitly gave listeners access to the intended bracketing structures, and the training material as well as the questions made it clear that this was a pairing disambiguation task, where the three items would be paired in a 2-1 or 1-2 structure. The emphasis, in other words, was on the conjunct pairs rather than on the meaning of the structure as a whole.

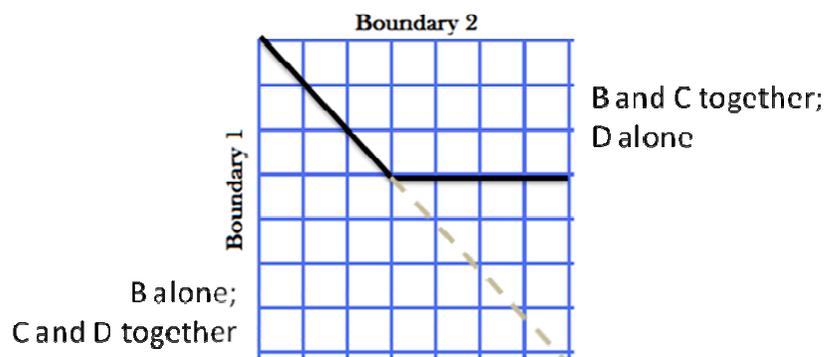


Figure 21: Predicted Contour Line for First Boundary Conditional Bias

In this case, a large boundary after the First constituent is informative, as it is consistent with a 1-2 pairing structure. Participants can therefore already begin

constructing the item structure at that point (with an early break structure), and this would lower the probability of a late-break interpretation for these items, resulting in what we call a First boundary Conditional Bias.

With this processing shortcut available, we would now predict that items in which listeners had direct access to the pairing structure would display a First Boundary Conditional Bias, whereas in other structures the Conditional Bias would still be dictated by the meaning-related processing capabilities of the sentence.

The opacity of the non-pairing items will become clearer as we discuss sentence-completion experiment paradigms and more complex structures in the following chapters. Whether the Formula item—where the answer choices were differently bracketed versions of the string—is to be defined as a forced-pairing choice or not is still up for debate. The results would suggest that speakers do not consider it a clear pairing choice that can be disambiguated at the first pause, and it is possible that as linguists we may be more sensitive to hierarchical branching structures than naïve language users, although more tests should clearly be run to verify this claim.

One further interesting point to note is that lengthening the boundary sizes for item *B and C and D* actually appears to have reduced the effect of the Conditional Bias, which is unexpected if the Conditional Bias does indeed depend on boundary strength. Unfortunately, a comparison with the *Rose and Steve and Kim* data is difficult, considering that the results for the short boundary intervals for that item were not very strong; but the long boundary intervals do show a strong Conditional Bias effect.

STUDY THREE: MODIFIED CONJUNCTIONS

3.0 Introduction

Having examined the behavior of simple conjunctions, in which all three constituents are of equal size and importance, we decided that the next step would be to consider modified conjunctions, or structures in which the first two constituents are still comparable in size and import, and the third is a modifier that can be interpreted as attaching either high, describing both of the conjuncts, or low, thus modifying only the second element.

Although these are structurally different, we predict that the listeners' processing strategy would attempt to disambiguate the forced-pairing task in a manner similar to the items in the previous study, thus forcing First-boundary Conditional Bias. The results are consistent with this model, although strong Overall Bias effects prevent a close investigation into the precise phonetic properties of the Conditional Bias.

3.1 Method

The items in this section represent a first foray into more lifelike syntactic structures and naturally occurring ambiguities, but the method employed to create, manipulate, and test these items was identical to the one in the previous section.

3.1.1 *Stimuli*

The stimuli used in this section were taken from Clifton, Carlson and Frazier (2002), where they were tested and it was found that they could be successfully disambiguated through the use of differently sized prosodic boundaries.

8. a. Professional dancers | and skaters | with national awards
b. American farmers | and workers | with no health benefits

c. Five-star chefs | and wine-tasters | with their own tv show

Note that the two conjoined constituents are not precisely identical in size, the first being modified by an adjective whose scope could also be interpreted as ambiguous. Clifton, Carlson and Frazier deliberately inserted this extra word to unbalance the sentence, since according to Frazier et Al's (1984) findings (in which speakers were said to prefer boundaries that create balanced rhythmic groups), one could otherwise predict a bias towards the second boundary position, i.e. *dancers and skaters | with national awards*. The insertion of this adjective would allow for less-imbalanced groupings and would hopefully attenuate that source of bias, although it is something to keep in mind while analyzing the results.

Items (8a) and (8b) were both manipulated over 7 intervals of 30 ms each for both boundaries (divided in half between pause and duration cues), to yield a total of 49 sound tokens. Item (8c) was instead manipulated over 4 intervals of 60 ms each, with the cues divided in a 3:1 ratio between pause and duration, for a total of 16 tokens, to confirm with broad strokes that the behavior of the two other items can be extrapolated to larger phonetic durations.

All items were prepared following the methodology described in the previous chapters: a single base file was synthesized and manipulated to remove boundary and pitch information, and it was then fed through a Praat script that generated a matrix of sound files varying pause and duration increments at the two selected boundary locations.

3.1.2 Subjects

Participants were recruited from the Cornell undergraduate community and were compensated with \$5 or one extra credit point for their undergraduate psychology class. Seven subjects were run on item (8a), five on (8b), and six on (8c).

3.1.3 Experiment Setup

Subjects participating in this experiment were made aware of the ambiguity present in these sentences during a training session before the experiment, in which they were asked to read a sentence containing the modified NP items listed above, where the intended grouping of constituents was conveyed by inverting the two constituent groups minus the initial adjective, and with the addition of modifiers *only* and *both* further reinforcing the intended meaning. as in:

9. a. The TV show invited professional dancers, and skaters with national awards... *In other words, it invited only skaters with national awards, and dancers.*
- b. The TV show invited professional dancers and skaters, with national awards... *In other words, both the skaters and dancers have national awards.*

The answer choices presented on screen were the inverted constituents only:

10. a. Skaters with national awards, and dancers
- b. Dancers and Skaters, both with national awards
11. a. Workers with no health benefits, and farmers
- b. Workers and farmers, both without health benefits
12. a. Wine tasters with their own TV shows, and chefs
- b. Chefs and wine tasters, both with their own TV shows

As before, participants were asked to select both a response as well as a confidence judgment for each token. The tokens were interspersed with filler items and presented under the same conditions as the other items in this thesis.

3.1.4 Conditional Bias Predictions

All three items in this section had an identical structure and task, and we would therefore expect that the variation across items, if there is any, would be restricted to Overall Bias of a lexical or contextual nature.

The processing scenario for this structure would suggest a Second Boundary Conditional Bias, but the forced-pairing task leads us to predict a First-boundary Conditional Bias, similar to those in the previous section. In the early break case, the modifier with-phrase attaches low to the second constituent, and the first part of the conjunction is held separately; whereas in the late-break case, the two conjoined items are paired closely, and the with-phrase attaches high and is held separately. In cases with a large First boundary, even though there is nothing to process at that point, subjects could already form an opinion about the pairing structure of the item, and the probability of the token resulting in a late-break interpretation is reduced.

3.2 Results

The same analyses were run on these data as for the previous items, and include both Univariate and Multivariate Analyses of Variance (ANOVA and MANOVA) tests, to assess the effects of possible independent variables on the dependent variable, overall and within levels of other factors. As before, significance was set at $\alpha = 0.05$, but results just shy of this mark are also reported and flagged for further testing and analysis.

3.2.1 Dancers and Skaters

The first item (8a) presented strongly significant First and Second boundary effects ($F = 9.575$ and $p < 0.001$ for First; and $F = 5.217$ and $p > 0.001$ for Second), with significant interaction between factors ($F = 1.568$, $p = 0.024$). The results of a

MANOVA analysis show significance at First within Second (0, 1, 3, 4, 5) and Second within First (2, 3, 4, 6) with $p > 0.05$. Figure 22 shows the distribution of responses, as well as the contour line demarcating the different areas of each interpretation.

The Mean Wresponse score for token (0,0) was 1.71, indicating a strong bias in the base sentence towards a late break interpretation (high attachment of the modifying phrase). This reinforces the impressionistic observation that the contour line in Figure 22 divides the Distribution Plot area giving more ground to the late break interpretation (above the line), than to the early break one (below the line).

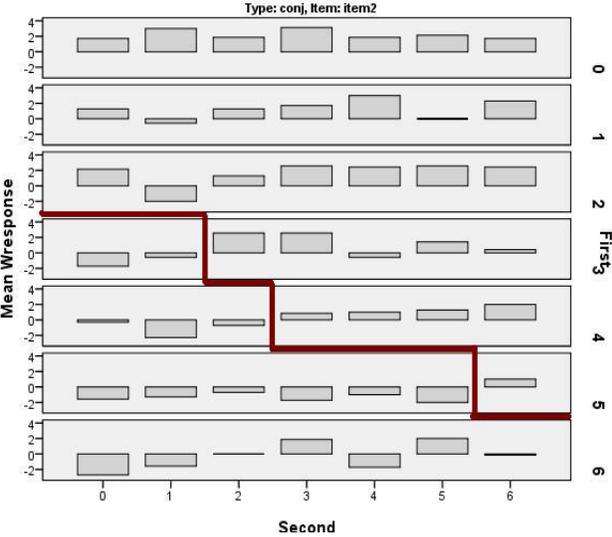


Figure 22: Distribution Plot for item "Dancers and Skaters"

The Difference variable is also significant in affecting the distribution of the Wresponse variable ($F = 5.530, p < 0.001$), but a MANOVA analysis shows that there is statistically significant variance ($p < 0.05$) for all Difference levels except -6, -5, -4, -2 and 4. This can be also seen in Figure 23, where the variation in results is limited to Difference levels 0, 1, 2, and 3, and there is otherwise a very clean and clear cut distribution of interpretations.

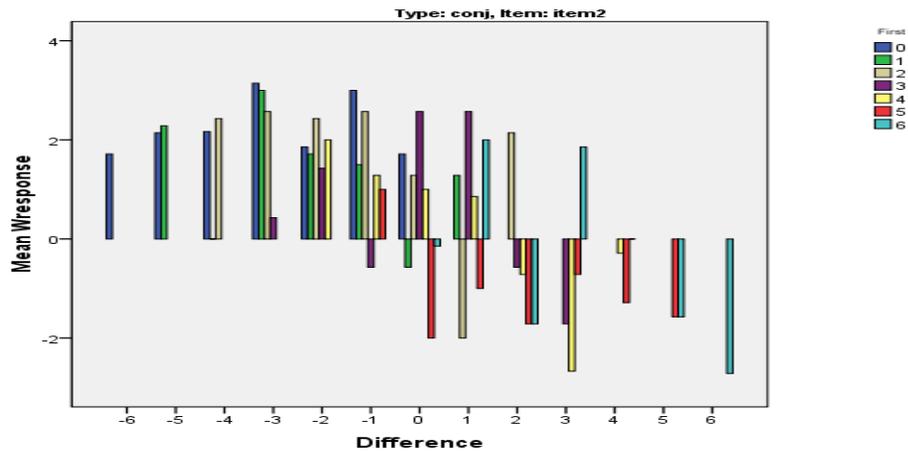


Figure 23: Difference Bar-chart for item "Dancers and Skaters"

3.2.2 *Farmers and Workers*

The same manipulations applied to a similar structure however produced very different results: for item (8b), neither the First nor Second boundaries are significant ($p > 0.350$), nor is there interaction between factors ($p = 0.120$). A quick look at Figure 24 quickly explains the lack of statistical significance as a by-product of the relatively low confidence subjects had in their interpretation of the sentence, particularly of the late break version. This could be due to any of a number of factors, from issues with the lexical or semantic content of the sentence, to problems understanding the synthetic speech, to difficulties with the way in which the item was presented.

It is therefore not surprising that the score for token (0,0) is a mere 0.66, even though the shift in the contour line would suggest a strong bias value in favor of the early break interpretation (which would mean a large negative number). The amount of noise, and in particular the strong variation in scores or magnitude of confidence, that is present across neighboring scores throughout the entire plot suggests that this

item might have been confusing for listeners, and they did not exclusively consider the prosodic cues when selecting the meaning and confidence rating for this item.

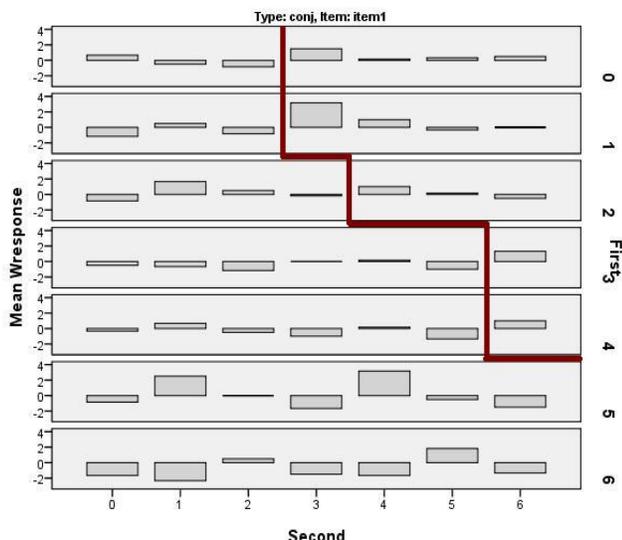


Figure 24: Distribution Plot for item "Farmers and Workers"

Figure 25 Figure 25 more closely, it appears that levels -2 and 1 actually present the most confusing picture of all, with alternating judgments which are difficult to explain linguistically.

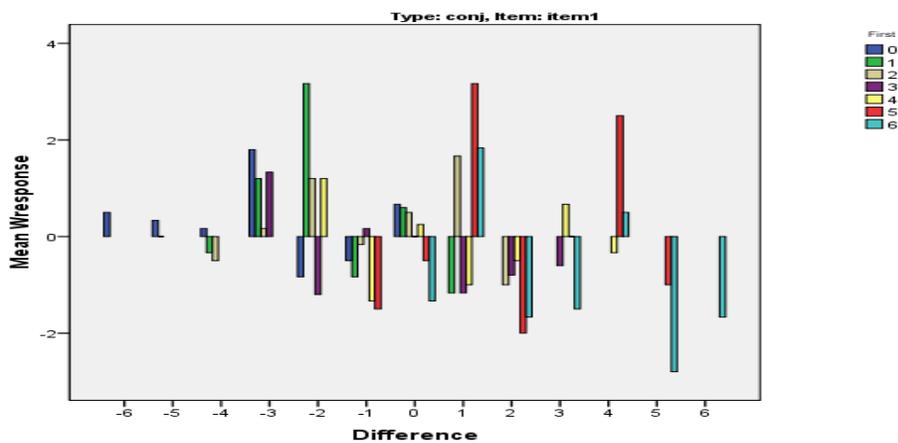


Figure 25: Difference Bar-chart for item "Farmers and Workers"

3.2.3 Chefs and Wine-Tasters

To further test the findings, item (8c) was tested on longer increments, ranging up to 90 ms of lengthening and 270 ms of increased duration total. As both boundaries already showed good results at shorter increments, the number of overall steps was reduced from 6 to 3, making each increment an increase of 30 ms lengthening, 90 ms pause.

With these extended intervals, significant differences in the distribution of the Wresponse variable are achieved mostly by manipulations of the First boundary, which is almost significant ($F = 2.612$, $p = 0.062$); while the second boundary is clearly not statistically significant in the distribution of the Wresponse scores ($F = 1.788$, $p = 0.162$). There is also no interaction between factors ($F = 0.525$, $p = 0.849$). The mean Wresponse score for token (0,0) was -0.25 , which under our assumptions does not indicate the presence of a strong bias in the base file.

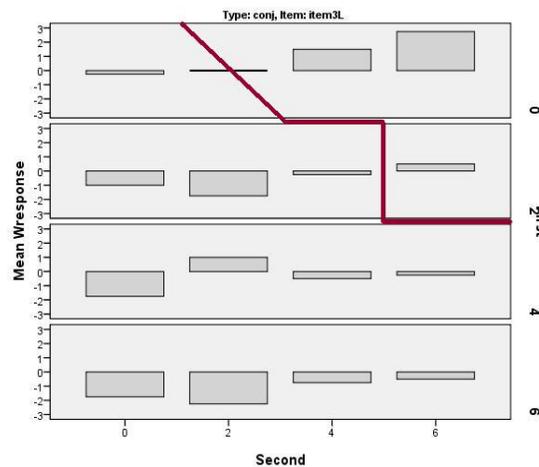


Figure 26: Distribution Plot for item "Chefs and Wine-tasters"

The Difference factor is significant ($F = 5.870$, $p < 0.001$) in affecting the distribution of the Wresponse variable, and MANOVA tests, as well as Figure 27 show that there is no significant effect of the First/Second factors.

Figure 26 and Figure 27 combined demonstrate how the distribution of responses for this item are only determined by the phonetic difference between boundary sizes, with a slight Overall Bias shift favoring the early break interpretation. In Figure 26, the contour line runs parallel with the Difference levels (which run from the top left to the bottom right corner), and in Figure 27, we can see that there is no variability within levels of the Difference factor, and the cross over point between late and early break interpretations occurs neatly between levels.

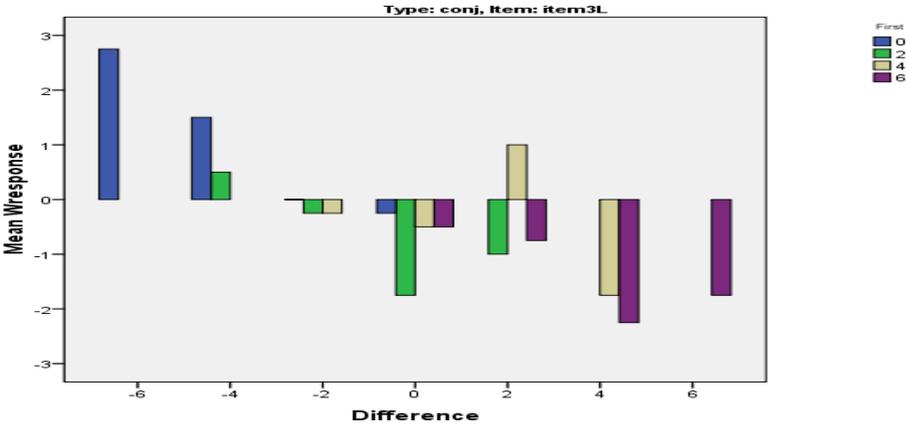


Figure 27: Difference Bar-chart for item "Chefs and Wine-tasters"

3.3 Discussion

There is unfortunately a lot of variability across the three items with respect to the statistical significance of the First and Second boundary factors. While both were significant for the first item *Dancers and Skaters*, neither the First nor Second are significant for the second item *Farmers and Workers*, while only the First and not the Second are significant on the longer intervals in *Chefs and Wine-tasters*. This in turn

affects the MANOVA scores both within First and Second factor levels, as well as within levels of the Difference factor, and may cause variability that is actually present not to show up in the results.

However, an examination of the Wresponse distribution graphs across items shows that the lack of statistical significance does not imply lack of variation and contrast in interpretations. All three items appear to have strong Overall Biases, which shift the contour diagonal equally across the First-by-Second plot. *Dancers and Skaters* has a slight Overall Bias which favors the late break interpretation, while *Farmers and Workers* and *Chefs and Wine-tasters* seem to prefer the early break option.

Only item (8a), *Dancers and Skaters*, shows some effects of the Conditional Bias, which is triggered by a First boundary of level 4 (120 ms) or above. Due to the strong Overall Bias affecting the other two items, items with a First boundary of that size or above are already predicted to be judged as having an early break: it is therefore impossible to tell whether the Conditional Bias is applying redundantly or not.

STUDY FOUR: PARTICLE VERBS

4.0 Introduction

In this section, we analyze another structure that has often been considered a source of potential attachment ambiguity, and was extensively analyzed in Price et Al (1991) but was not treated by Clifton, Carlson and Frazier perhaps due to the syntactic constraints they placed on the domain in which the hypothesis applied.

This is again one step more complicated than the previous structure: in these items, both the first and last constituent are now longer and can be internally complex, but we attempted to conserve the simplicity of the middle constituent (the preposition/verbal particle) which is still monosyllabic, monomorphemic, has a simple phonetic/syllabic structure, and has limited internal semantic content precisely due to its mono-morphemic nature. The combination of this particle with the other two constituents is however more complex than the simple association or scope relations that were present in previous items, and for this reason it was necessary to modify the experimental task as will be described in section 4.1.3.

4.1 Method

In this section we finally consider realistic ambiguities found in full sentences, and this required a number of changes to be implemented in the experiment procedure, described in detail below.

4.1.1 Stimuli

We approached this structure with the intention of recycling many of the Price et Al (1991) items in order to allow for results to be compared across experiments, but soon ran into issues of grammaticality or strong bias, both at the planning stage and often from subjects' comments during the training sessions (in which they were asked

to report to the experimenter if they felt certain items were ungrammatical or too hard to imagine). The final five items, designed to have the same structure as the Price et Al originals, were the following:

13. a. The tourist checked | in | the bags.
- b. The student dropped | off | the table.
- c. The Vikings won | over | their enemies.
- d. The tires may wear | down | the road.
- e. The engineers looked | up | the elevator shaft.

All items were tested in the standard 7 by 7 matrix with short intervals (30 ms divided equally between pause and duration), and items (13a) and (13c) were also tested in a 4 by 4 matrix with long intervals, of 120 ms (in a 1:3 ratio between lengthening and pause insertion).

4.1.2 Subjects

Native speakers of American English were recruited from the Cornell undergraduate population, and paid \$5 for participation in this experiment. Five subjects were run on the short version of item (13a), and four on the long version; fifteen subjects were run on item (13b), and thirteen on (13c) in the short version, and a further four on the long version. Item (13d) was run on four subjects, and (13e) was run on five.

4.1.3 Experimental Setup

The setup of the experiment differs slightly from previous ones, in that the ambiguous sections are now entire sentences, and more than simple punctuation is required to successfully disambiguate the different meanings. To facilitate the comprehension and disambiguation of the two meanings, we decided to alter the task

so that it would be a sentence completion choice. The answer choices are as follows, and as before, the participants were exposed to the two possible interpretations in a written training session.

14. The tourist checked in the bags...
 - a. ... and proceeded to the departure gate.
 - b. ... to see if he had forgotten his passport.
15. The student dropped off the table...
 - a. ... and got tipped for the delivery
 - b. ... and passed out, drunk
16. The Vikings won over their enemies...
 - a. ... by peaceful trading and persuasion
 - b. ... and annihilated them in a bloody war.
17. The tires may wear down the road...
 - a. ... because their reinforced core will erode the asphalt
 - b. ... because they're not well constructed and will wear through quickly.
18. The engineers looked up the elevator shaft...
 - a. ... in the blueprints to review the measurements.
 - b. ... to check for dangling cables.

4.1.4 Conditional Bias Predictions

In this section we can finally test the predictions of our processing hypothesis on real language data, with the real meaning calculations it would entail. The process is exactly the same as before: take the item *The student dropped off the table* as an example, where the pause locations here occur before and after the preposition *off*. At the first pause location, it would already be possible to process the preceding constituents and calculate the meaning of the phrase as an intransitive, where the

student is the patient of the verb dropped. Given enough time for the calculation of this meaning, this would decrease the likelihood of a late break interpretation, consistent with a First Boundary Conditional Bias effect.

A long pause at the Second boundary location is void, since there has already been a previous Point of Disambiguation in the sentence, which was responsible for triggering the appropriate Conditional Bias.

4.2 Results

As in other experiments, ANOVA and MANOVAs were carried out on the Wresponse score distributions to test the significance (set at $\alpha = 0.05$) of the factors First, Second and Difference, both overall and within each other.

4.2.1 Check in

The first item showed significance for both factors, with $F = 3.247$ and $p = 0.005$ for the first boundary, and $F = 2.268$ and $p = 0.039$ for the second. The interaction of the two factors was not significant ($F = 0.839$ and $p = 0.729$). MANOVA tests showed significance only for one factor level each in First within Second and Second within First. However, the contour line shown in Figure 28 clearly shows that both First and Second boundary influence the distribution of the Wresponse score, and in fact, the two have virtually equal weight, as the line remains diagonal throughout.

The Mean Wresponse score for token (0,0) is a mere -0.2, supporting the view that there is no strong Overall Bias present for this item, a finding reflected by the contour line in Figure 28, that flanks the token (0,0) and continues at almost a perfect diagonal throughout the entire distribution plot.

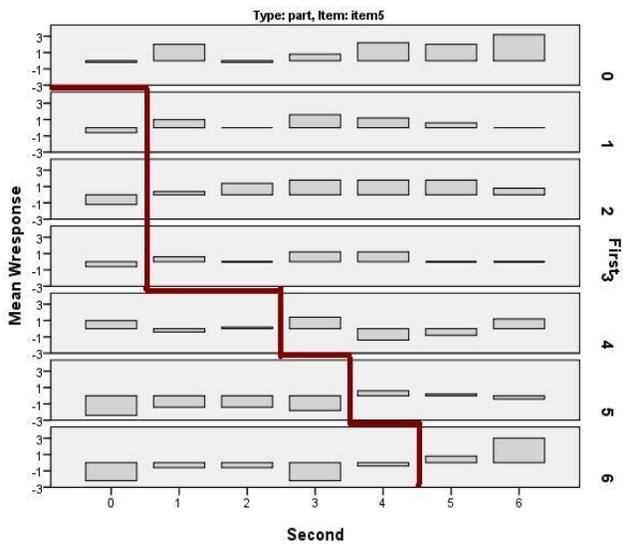


Figure 28: Distribution Plot for item "Check In" (short)

The Difference variable is also significant, with $F = 2.794$ and $p = 0.001$; MANOVAs of the First/Second boundary effects within the levels of Difference are significant only for Difference levels of 3, 5 and 6. Figure 29 displays a very clean distribution of scores, with no systematic variations (the alternations appear to be only due to random noise in the responses, and a switch between overall late and early break interpretations between Difference levels 1 and 2 or 3).

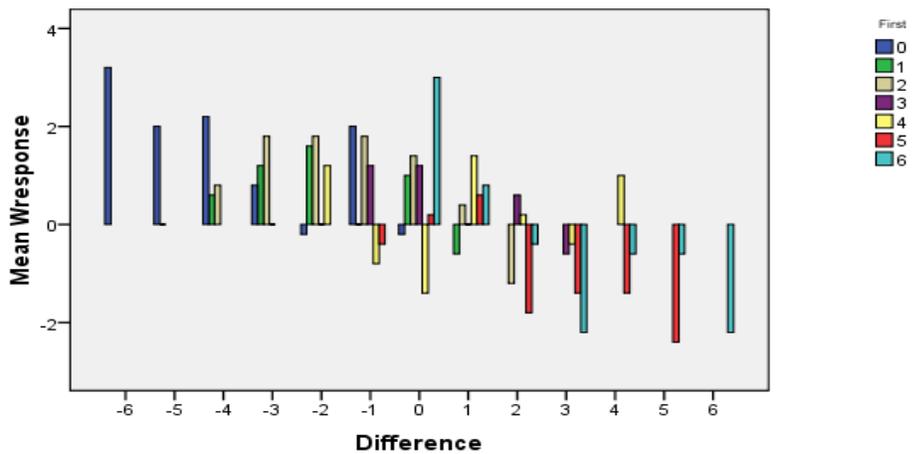


Figure 29: Difference Bar-chart for Item "Check In" (short)

The same item was also run on longer intervals (4 levels separated by 120 ms, divided in a 3:1 ratio between pause and duration cues); and for this the First boundary was significant ($F = 5.526$, $p = 0.002$), but not the second ($F = 1.121$, $p = 0.350$), nor the interaction between boundaries ($F = 0.972$, $p = 0.350$). MANOVA tests show significance of the First boundary within levels 0 and 2 of the Second ($p < 0.05$), but nowhere else. Not surprisingly, the variable Difference is also not significant ($F = 1.162$, $p = 0.340$).

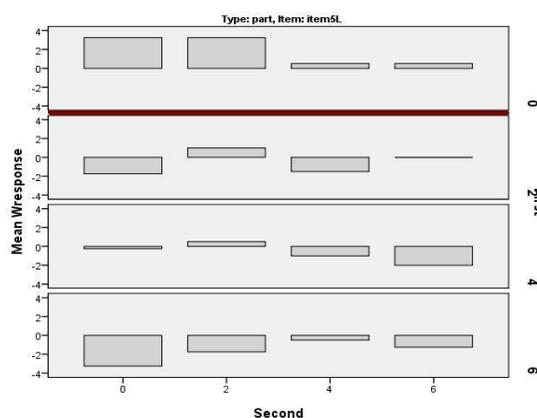


Figure 30: Distribution Plot for item "Check In" (long)

The Mean Wresponse score for token (0,0) is a startling 3.25, revealing extremely strong bias towards a late break interpretation—particularly when compared to the same item, in the shorter-grid version, which received a mere -0.2 .

Figure 31 shows a very clear switch in interpretations between levels 0 and 2 (0 and 120 ms of total boundary duration), and this is echoed in the systematic contrast between the first bar of Difference levels -6 through 0, which has late break scores, and the other boundary levels with early break interpretations.

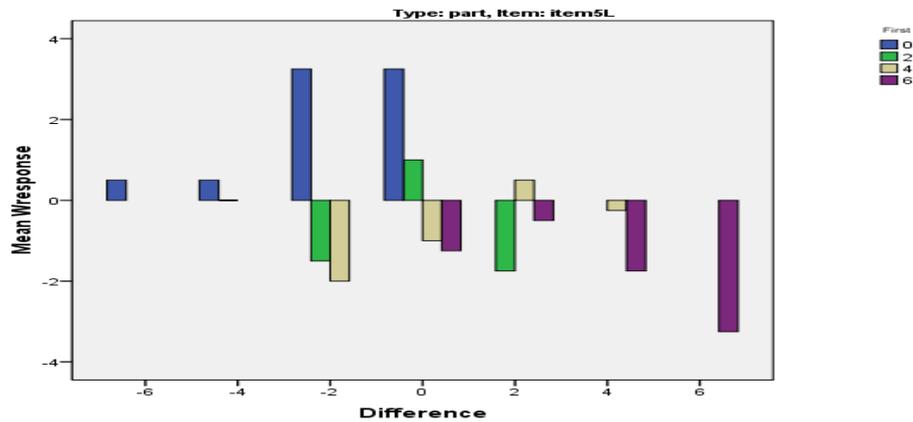


Figure 31: Difference Bar-chart for item "Check In" (long)

The short-interval item, whose results are described in Figure 28 and Figure 29, was also modified with First boundary levels in excess of 120 ms (the maximum lengthening was 180 ms)¹³, and as such the Conditional Bias effect should have been visible around factor level 4. This result is very surprising, and warrants further investigation, especially in light of the fact that other items' results (see section 4.2.3 for example) are consistent across multiple iterations of the same task.

4.2.2 Drop off

This item showed significance for both First ($F = 15.533$, $p < 0.001$) and Second ($F = 2.274$ and $p = 0.35$) boundaries, with no interaction of factors ($F = 0.730$, $p = 0.879$). The MANOVA tests show significance of First boundary in Second at all levels except 5 ($p < 0.02$), but never of the Second boundary within levels of the First. This asymmetry between the effects of the First and Second boundary can be clearly

¹³ The only exception would be if listeners were sensitive only to the pure pause duration, which would be of 90 ms for the Conditional Bias shown in the 4 by 4 matrix—this would correspond to level 6 of the shorter-interval 7 by 7 matrix, so leaving room for some variation the Conditional Bias effect might actually have occurred “off the matrix”, but this seems unlikely.

observed in Figure 32, where the contour line flattens out, clearly indicating a First boundary Conditional Bias between levels 3 and 4.

The Mean Wresponse score for the token (0,0) was of 0.73, which suggests that there is no strong Overall Bias present in this item. The contour line in Figure 32 similarly suggests the presence of at most a slight Overall Bias towards a late break interpretation, but even Figure 33 does not show clear evidence of this change (cf. Difference = 1 and 2), due to the low magnitude, or confidence ratings, of these tokens.

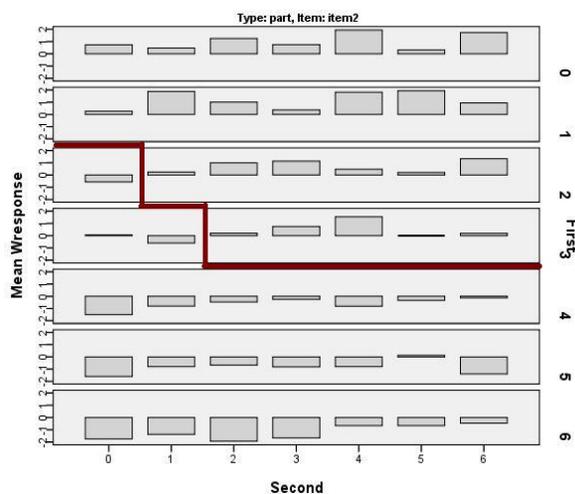


Figure 32: Distribution Plot for Item "Drop Off"

The difference variable is significant at $p < 0.001$ ($F = 7.309$), but MANOVA tests, as well as Figure 33 on the following page, show that there is strong variability within levels of the Difference boundary and this is statistically significant at Difference = -6, -4, -1, 0, 3, 4, 5 and 6. Figure 33 however clearly shows that there is systematic variability of the First/Second boundary factors within levels -1, 0 and 1, while the more extreme values of the Difference factor in either direction show consistent responses across all levels of the First/Second boundary factors, indicating a neat distribution of responses on either side of the contour line.

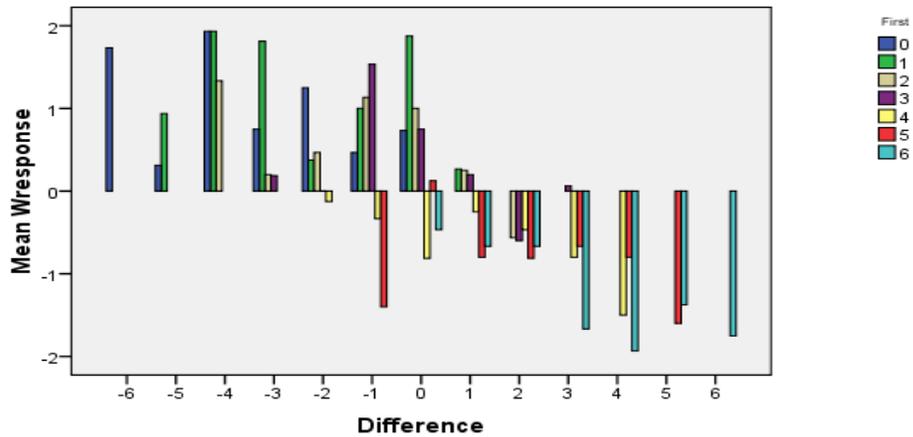


Figure 33: Difference Bar-chart for item "Drop Off"

The distribution of Wresponse data is consistent with the interaction of a slight Overall Bias which favors the late break interpretation, and a Conditional Bias which is triggered by First boundaries that are size 4 (120 ms) or larger.

4.2.3 Win Over

This item displays a strong significance of the first boundary at $F = 10.674$ and $p < 0.001$, but the second boundary factor is not at all significant ($F = 0.664$, $p = 0.679$), nor was the interaction between factors ($F = 0.699$, $p = 0.908$). MANOVA tests show that the First boundary is or approaches statistical significance within all levels of the Second boundary except Second = 6 ($p < 0.07$, $F > 1.95$); but the Second boundary is at no point significant within levels of the first. Figure 34 clearly shows that this is due to a radical change of interpretation between First boundary levels 1 and 2, across virtually all levels of the Second boundary, consistent with a very strong First boundary Conditional Bias situation.

The Mean Wresponse variable for token (0,0) is 0.66, which again does not qualify as representative of a strong Overall bias, by our definition, and this is

reflected in the Distribution Plot by the short diagonal portion of the Contour Line which flanks the point (0,0).

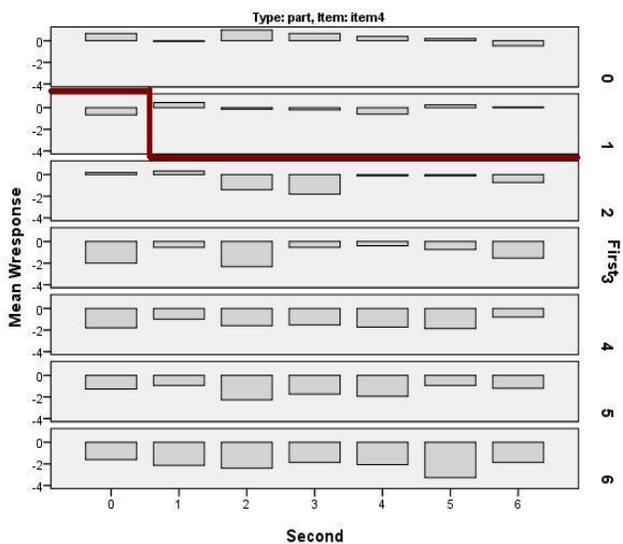


Figure 34: Distribution Plot for Item "Win Over" (short)

The Difference variable displays statistical significance ($p = 0.001$, $F = 2.846$), and MANOVA tests show that there is a statistically significant effect of the First/Second boundaries within levels 0 and 1 of Difference only, and not elsewhere ($p > 0.1$). The systematic variation which somehow is not captured by the statistical tests is between the first two bars and all the other bars within the clusters corresponding to Difference levels -6 through 0.

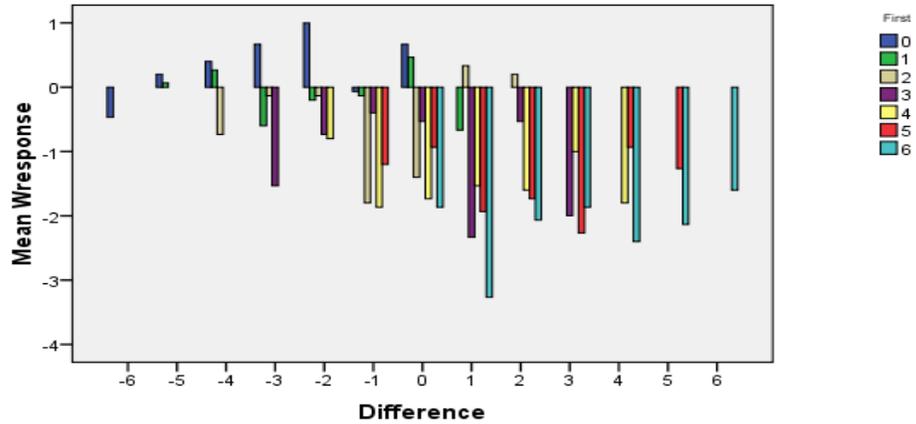


Figure 35: Difference Bar-chart for Item "Win Over" (short)

The same item was also run on long intervals (4 steps of 120 ms divided in a 3:1 ratio between pause and duration cues), and the results are remarkably similar to those presented by the longer version of item 13b *check in*. The First boundary is here strongly significant ($F = 6.400$, $p = 0.001$), but the Second boundary isn't ($F = 2.175$, $p = 0.103$), nor is the interaction between factors ($F = 0.637$, $p = 0.760$).

The factor Difference is barely significant, with $p = 0.036$ ($F = 2.441$), and MANOVAs show that there is a strongly significant effect of the First/Second boundaries within the levels of Difference except for Difference = -2, and 4.

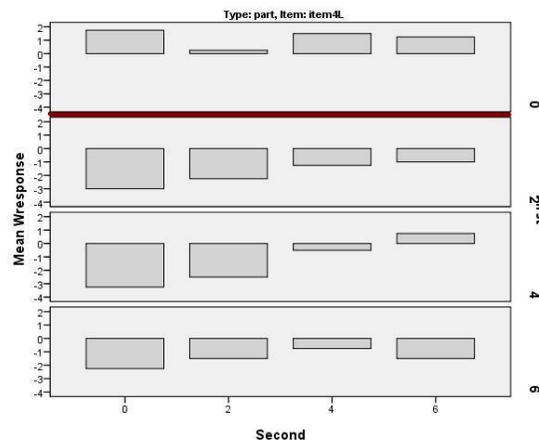


Figure 36: Distribution Plot for Item "Win Over" (long)

Once again, we see that in this longer variation of the item, the Mean Wresponse token (0,0) score rises to 1.75, which indicates a strong bias towards late break interpretation in the base file—which is startling, considering that the same exact token received a score of just 0.66 in the short version of this item. However, it is impossible to see whether this really corresponds to a strong Overall Bias, as the extremely strong First boundary Conditional Bias effect draws a perfectly horizontal contour line through the Distribution Plot.

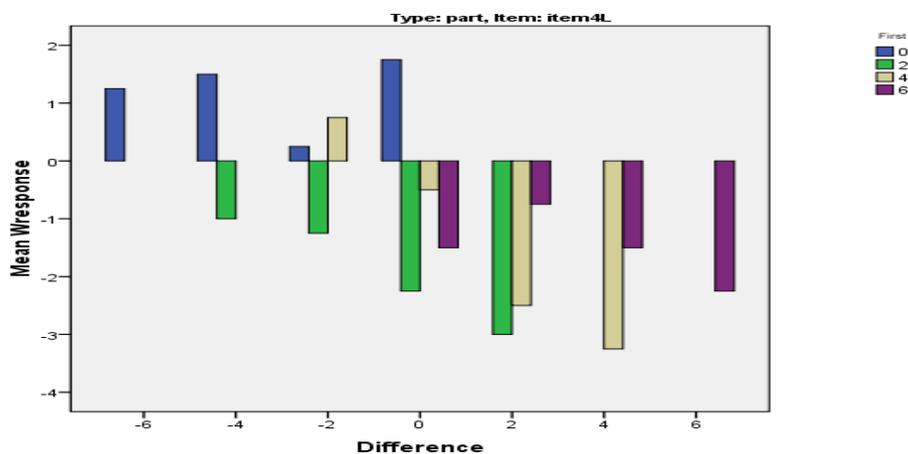


Figure 37: Difference Bar-chart for Item "Win Over" (long)

Figure 36 and Figure 37 display the same characteristics as those for the smaller-interval 7 by 7 matrix for tested on the same item, discussed on the previous page: the contour line in Figure 36 shows a clear case of First boundary Conditional Bias affecting all boundaries size 2 (120 ms) or above, and this effect is mirrored in the systematic variation of the first bar within the clusters of Difference levels -6 through 0 of Figure 37.

4.2.4 Wear Down

Item (13d) showed identical results: significance only for the First boundary factor ($F = 8.510$ and $p < 0.001$), but not for Second ($F = 0.311$ and $p = .931$) or interaction between factors ($F = 0.402$, $p = 0.999$). The MANOVA tests showed only limited effects of the First boundary within levels 3 and 5 of the Second boundary, and no significant effects elsewhere.

The Mean Wresponse score for token (0,0) for this item was of just 0.25, suggesting—as per our assumptions—that there is no strong Overall Bias affecting all tokens. This cannot however be confirmed graphically due to the overwhelming effect of the Conditional Bias, which draws a perfectly horizontal line through the Distribution Plot.

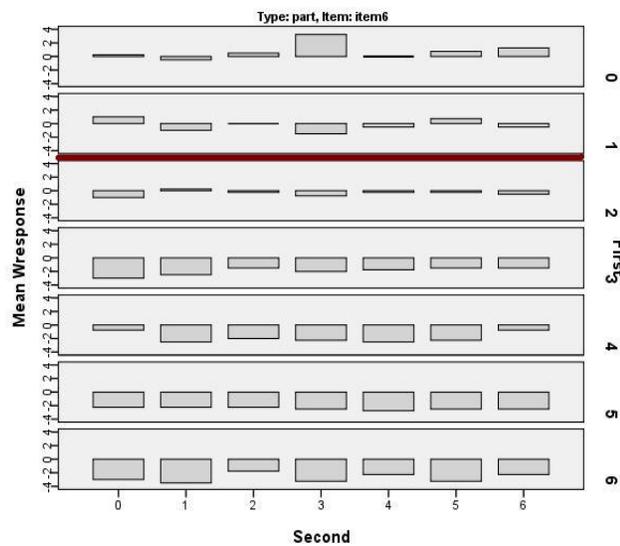


Figure 38: Distribution Plot for Item "Wear Down"

The Difference variable is strongly significant ($p = 0.002$, $F = 2.789$), and MANOVA tests show that there is no significant effect of the First/Second boundary

distribution within the levels of the Difference variable (except for Difference = -6 and -4).

Graphically, these results are shown in Figure 39 and 40, and as before show a very clear First Boundary Conditional Bias, where a boundary of level 2 (60 ms) or above automatically triggers early break interpretation, and this can also be seen in the systematic differences between the first two bars (vs. all other bars) of Difference levels -6 through 1.

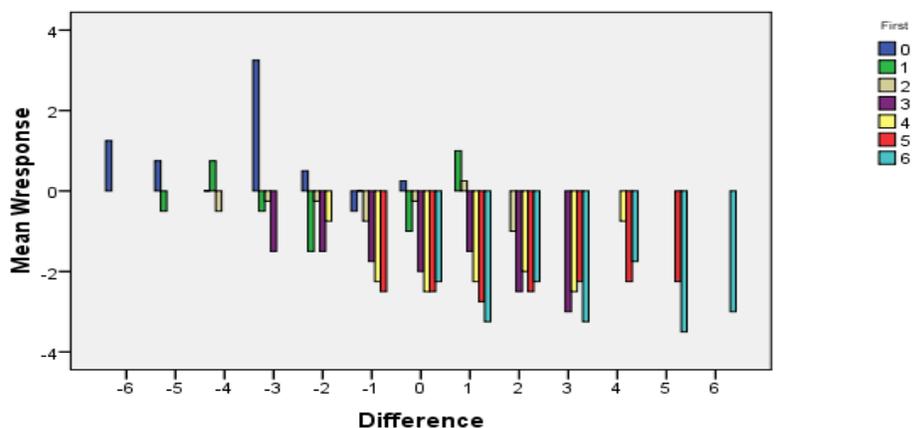


Figure 39: Difference Bar-chart for Item "Wear Down"

4.2.5 Look Up

The last item tested for this structure had such strong bias effects that it did not switch interpretations even for the most extreme prosodic boundary differences: that is, even a token with 180 ms total boundary duration (90 ms pause, 90 ms lengthening) after *up*, and none after *looked*, in the structure *The engineers looked up the elevator shaft*, was unable to consistently induce late-break interpretations. Not surprisingly, neither the First, nor Second, nor Difference factors are significant in affecting the distribution of the Wresponse variable ($p > 0.200$).

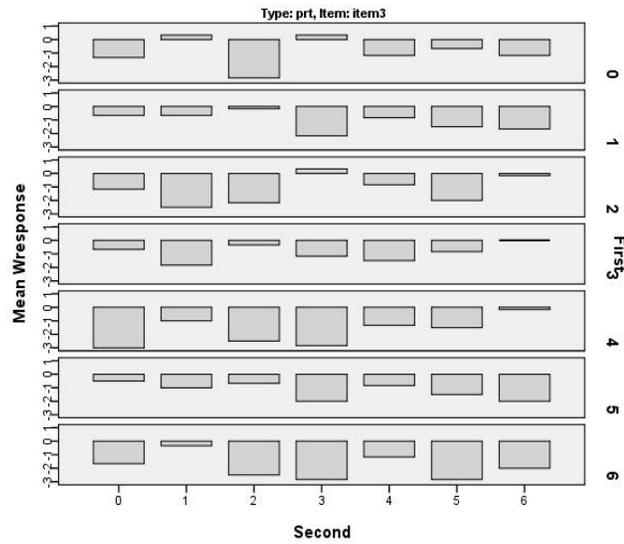


Figure 40: Distribution Plot for Item "Look Up"

The Mean Wresponse score for token (0,0) was of -1.33, reflecting the fact that there is strong bias in the base file towards an early break interpretation of this structure. One possible cause for this could be a lexical bias against the particle verb form of *look up*, possibly dictated by the directionality component of the PP (*up the elevator shaft*), which either decreases or completely zeroes out the probability of a late break (particle verb) interpretation.

Alternatively, this could be caused by acoustic properties of the two boundaries, where pause insertions at the second boundary location (between *up_the*) are always considered part of either the preceding or following stop closure or release, whereas the pause insertions at the second boundary location (between *look_up*) are always considered at least partially to be a boundary.

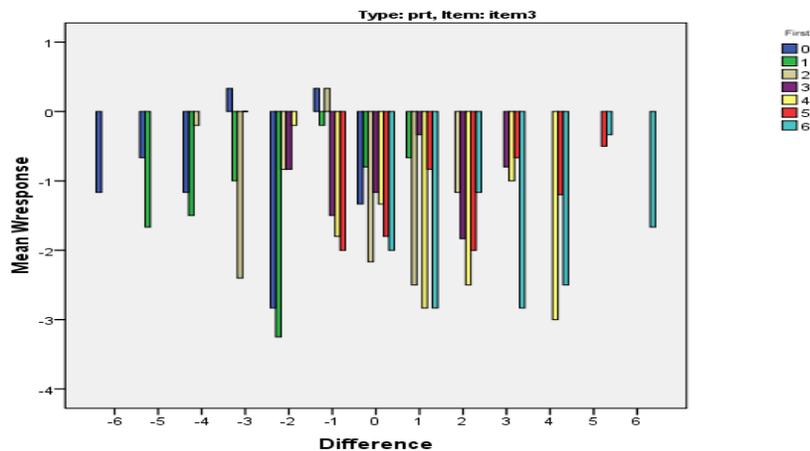


Figure 41: Difference Bar-chart for Item "Look Up"

This item, while not very interesting for our purposes, is included to show that even though subjects can be fully aware of the two available meanings of an ambiguous sentence, non-prosodically-induced bias can be strong enough to counteract even very strong prosodic cues designed to induce the opposite interpretation (which function perfectly well in other contexts)¹⁴.

4.3 Discussion

This study was the first to contain actual complete sentence ambiguities, and we were expecting strong variation across items to reflect lexical differences, but (with the exception of the unruly short-interval *Check in* data), the results were virtually identical across conditions.

14 Note that this is not a case in which the particle-verb situation was deemed ungrammatical or strongly dispreferred by subjects, as subjects did not report any problems when exposed to this item in the training session. Ungrammatical items included a variation on Price et Al's structure: *Margaret rolled over the carpet*, whose particle verb reading was available only as *Margaret rolled the carpet over* for most subjects.

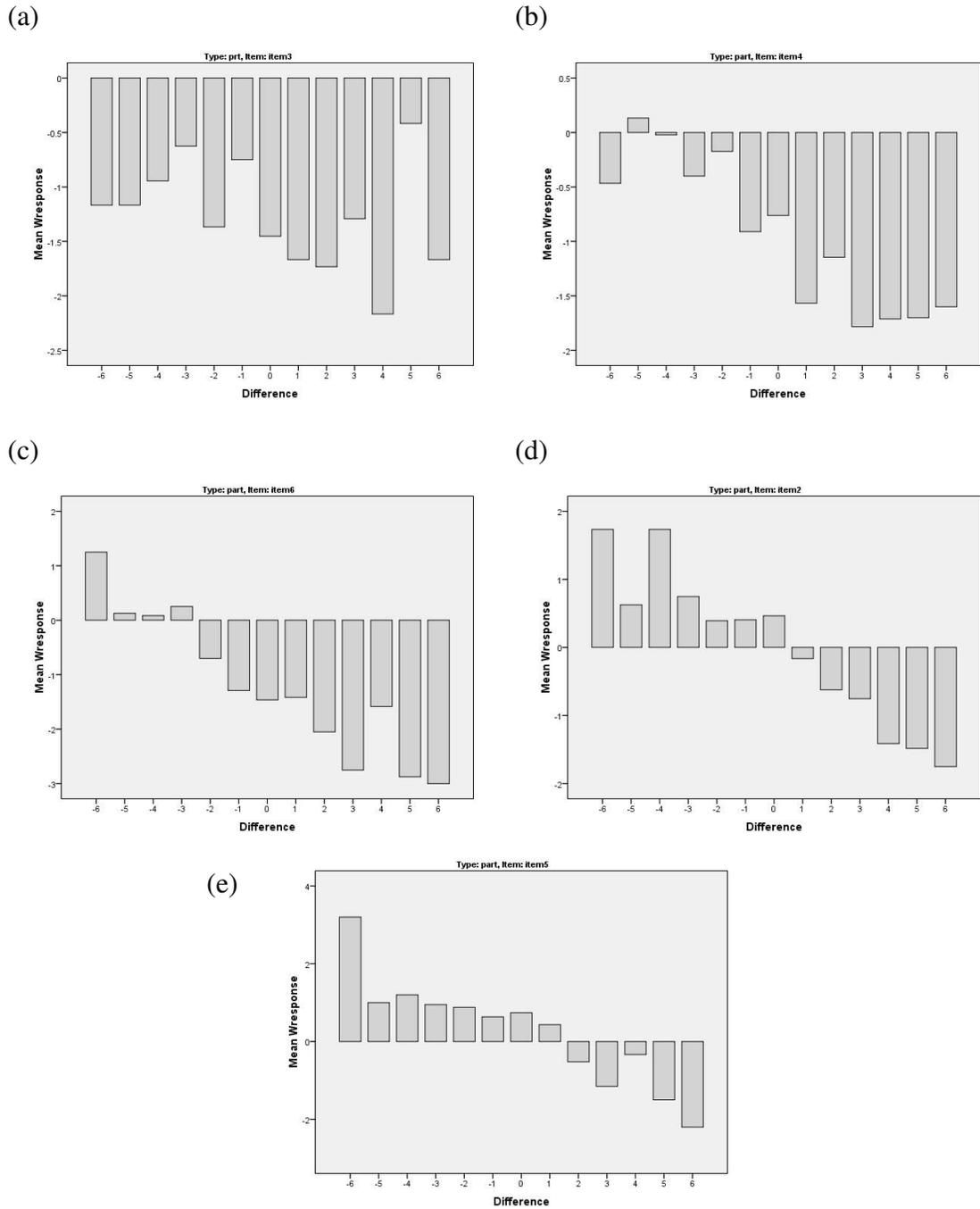


Figure 42: Comparative Difference Bar-charts for Particle Verb items
 “Look up”, “Win Over”, “Wear Down”, “Drop Off” and “Check In”

All items displayed extremely strongly significant First boundary effects, with varying degrees of Second boundary significance, and all items displayed behavior consistent with a First boundary Conditional Bias, according to which First boundaries that are 60-120 ms long would affect subjects' processing of the sentence in such a way that the probability of a late break interpretation is significantly diminished.

This isn't to say that there isn't variability across items: rather, this variability appears to be captured solely by the Overall bias, which shifts the contour (the line of perceptual equality) to different levels of the Difference boundary, as the item goes from early- to late-break bias. Figure 42 shows the Mean Wresponse value for each level of the Difference variable, and the progression from the extremely biased *look up* (a), which doesn't even have a crossing point within the tested domain, to the marginally biased *win over* (b) and *wear down* (c) items, which show a slight bias towards the negative side of the Difference scale, to the balanced *check in* (e) and *drop off* (d) items.

These results are highly consistent with our processing model, which predicts a consistent source of Conditional Bias across items with the same structure (and experimental task), but allows for a range of variation between items depending on the lexical or phonetic properties of each, which can be accounted for and described by Overall Bias.

However, even Overall Bias appears to be sensitive to the experimental context to some degree, as in this chapter, the two items tested in multiple grid-sizes received strongly differing ratings for the token (0,0), whose phonetic properties should be identical across these items¹⁵.

¹⁵ Technically, the two tokens of (0,0) in such comparisons are distinct sound files, as they were each generated from independent runs of the automated manipulation script that created all the boundary levels. However, since they were generated from the same base sound file, under the same set of parameters, the two files are phonetically identical.

A close examination of the data reveals that this sharp contrast (between -0.2 and 3.75 for *Check In*; and between 0.66 and 1.75 *Win Over*) is not simply the result of noise or an outlier in the data, as both the low Overall Bias values for the short-interval items, as well as the larger values for the long-interval items, are mirrored in neighboring tokens. It is possible to graphically confirm the absence of Overall Bias in the short-interval tokens for both items, as the Contour Line displays its typical diagonal bordering the (0,0) token. Unfortunately, the interaction of a strong First Boundary Conditional Bias with an Overall Bias favoring the late break interpretation obfuscates the effect of the Overall Bias, as shown in Figure 43, making it impossible to verify whether this is truly an effect of Overall Bias, or whether this is a magnification of the few non-early break scores present in the item.

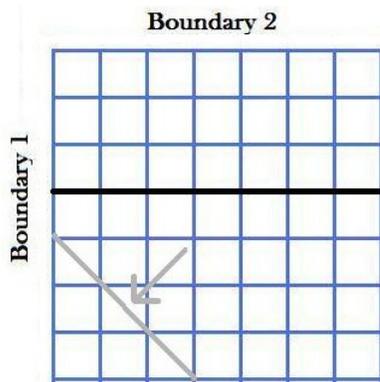


Figure 43: Predicted Interaction Effect of Strong Late Break Overall Bias and First Boundary Conditional Bias

However, since the sound files tested were exactly identical (generated separately but from the same base file and via the same manipulations), and the experimental task and training materials did not vary across the two items either, there appears to be no sensible source for such a large change in the Overall Bias of the item, and we are tempted to exclude that interpretation altogether.

The only difference between these two sound files can be retraced to the context in which they were presented, that is the matrix of 49 (or 25, in some cases) versions of the same sound file¹⁶. It is possible that being exposed to larger boundary sizes would have speakers recalibrate their sensitivity to pauses and give more extreme judgments for the same token when in the context of one or the other set of boundary extrema.

For these items, the larger phonetic range in the second version of these items would give stronger cues of early break interpretation given the First boundary Conditional Bias effects that are present. Coupled with a smaller matrix (only 25 items instead of 49), this means that there are fewer tokens in the distribution plot which get assigned the late break interpretation, and we propose that subjects unconsciously inflate the scores of these few tokens in order to somehow balance out the score distribution.

Unfortunately there are not many other items which were subject to the same contrast of longer and shorter intervals (two were described in the previous chapter, but did not show this score magnification), but we will be comparing the results for these items throughout the studies, in an effort to shed light on this interesting phenomenon.

¹⁶ This within-subjects experimental method was chosen to reduce variability in the Response scores across tokens, particularly given the relatively low number of subjects that we were able to run for each of the items within each structure.

STUDY FIVE: PREPOSITIONAL PHRASES

5.0 Introduction

We next discuss the instrumental-vs-modifier attachment of prepositional phrases, a classical example of attachment ambiguity discussed in the literature since Lehiste's work in the 1970s. Informal observations displayed a large amount of variability across items, presumably due to lexical or world knowledge bias, and for this reason we decided to test a larger number of distinct tokens, over fewer factor levels.

As predicted by our analysis of the processing points in the sentence, the results show a very strong Second boundary Conditional Bias, which is consistent in location and strength across different trials with the same item as well as across items. However, not all items do display Conditional Bias, suggesting that strong Overall Bias effects may interact with the Conditional Bias and its domain.

5.1 Method

In this experiment, we decided to test a larger number of items over smaller matrices, but the experimental setup otherwise remained identical to that presented in the previous chapter. Details are included below.

5.1.1 Stimuli

Item (19a) was first run on the standard 7 by 7 matrix over short intervals (30 ms each, divided in half between pause and duration cues), to compare it to all other items run in this thesis. However, the results were inconclusive, so it was decided to test it on the same 7 by 7 matrix with longer intervals (60 ms per interval, divided over a 3:1 ratio between pause and duration), and seeing that the results were extremely clear, the remaining items were tested on four by four matrix structure, over the same

long time series (360 ms max lengthening, divided 3:1 over pause and duration—120 ms per interval).

19. a. The girl tried to wake | the sleeper | with the teddy bear.
- b. The police attempted to follow | the felon | with the Rottweiler.
- c. The principal began to lecture | the student | with a worried expression
- d. The seamstress managed to ruin | the dress | with a bow.
- e. The enemy continued to batter | the fortress | with the cannon.
- f. The soldiers tried to locate | the rebels | with the attack plan.

The stimuli were chosen as the most neutral of a larger set of 40 items that had been constructed for this purpose—five non-naïve native speakers of English were asked to rate whether they found high or low attachment more natural, and the six items with the most neutral average judgment were selected for this study. The sentences were then synthesized and manipulated in the same manner as all the others in this thesis.

5.1.2 Subjects

Native speakers of American English were recruited from the Cornell undergraduate population, and compensated \$5 or one extra credit point for their time. All of the small matrix items (4 by 4) were run on eight subjects; while item (14a) in the 7 by 7 format and long intervals was run on six subjects, and the pilot item with the 7 by 7 and short intervals was run on 3 subjects.

5.1.3 Experiment Setup

The experiment setup remains unvaried from that used for particle verbs, with a written training session preceding the study, in which the experiment items were

interleaved with fillers and presented to subjects over headphones in a sound proof booth.

As before, subjects were asked to provide both a response and a confidence judgment for each token. The responses were possible sentence continuations, as shown below:

20. The girl tried to wake the sleeper with the teddy bear...
 - a. ... but the sleeper kept holding the teddy bear and snored on
 - b. ... but she soon decided that a tambourine was more effective than a stuffed animal
21. The police attempted to follow the felon with the Rottweiler...
 - a. ... because he seemed to be the most dangerous of the gangsters
 - b. ... because they didn't have any German Shepherds available on site
22. The principal began to lecture the student with a worried expression...
 - a. ...but the principal didn't take pity on the worried student and kept lecturing
 - b. ...but the principal's angry words didn't match his concerned look and nervous tone
23. The seamstress managed to ruin the dress with a bow...
 - a. ... but the dress with the embroidery was fortunately not damaged
 - b. ... because the addition of the bow completely ruined the gown's silhouette
24. The enemy continued to batter the fortress with the cannon...
 - a. ...because if they demolished the only armed fortress, the city would be defenseless
 - b. ...and the cannonballs were seriously damaging the city walls"
25. The soldiers tried to locate the rebels with the attack plan...

- a. ...to interrogate them and find out who gave the rebels the details for the attack
- b. ...but the soldiers' attack plan was flawed and they burst into an empty house

5.1.4 Conditional Bias Predictions

The location of the two boundaries for these items was chosen so that they would mirror the structure of the early and late breaks that have been used for other items. In pronouncing these items, however, you will most likely find a strong resistance to creating large boundaries in the First location, as this would be consistent with an intransitive reading of that portion of the sentence.

- 26. a. The girl tried to [[wake] [[the sleeper] [with the teddy bear]]]
- b. The girl tried to [[[wake] [the sleeper]] [with the teddy bear]]

This is exactly what is predicted to happen during processing: a large boundary at the First location is uninformative, as it would suggest that speakers build the structure with the subject as a patient—meaning *The girl tried to wake herself up*, and in some cases this intransitive reading might not even be available. As soon as the next constituent begins, though, it is obvious that the structure is transitive and has to be rebuilt.

A large Second boundary can be interpreted as a sign of a break between the Verbal cluster and the with-phrase, and would trigger high attachment of the prepositional phrase, giving it an instrumental reading. This means that the Second boundary location is the only tested Point of Disambiguation for these structures, and given enough time at this location, we would expect the presence of a Second boundary Conditional Bias that would lower the probability of an early-break (low attachment) reading.

5.2 Results

As in other experiments, ANOVA and MANOVAs were carried out on the Wresponse score distributions to test the significance (set at $\alpha = 0.05$) of the factors First, Second and Difference, both overall and within each other.

5.2.1 *Teddy bears*

This item was the most extensively tested, as it served as a pilot to gage the exact properties of the phonetic manipulations to be applied to the other items of the structure. It was first tested in using the standard phonetic manipulations, using is a 7 by 7 grid with each interval consisting of 30 ms of boundary increase, divided equally between pause insertion and pre-boundary lengthening.

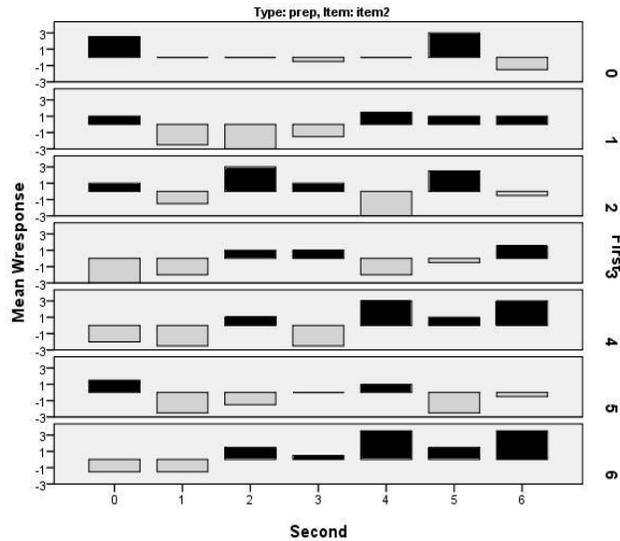


Figure 44: Distribution Plot for "With the Teddy Bear" (short)

The results for this item are shown in Figure 44, with the shaded squares representing tokens with a mean Wresponse score indicating late break interpretation, and the light squares representing tokens with a mean Wresponse score indicating

early break interpretation. The squares appear to alternate in a random pattern and cannot be divided into consistently early or late break areas. Furthermore, none of the variables are statistically significant ($p > 0.200$) and for this reason it was decided to retry the study using larger interval gaps between factor levels.

In the revised study, each interval consisted of an addition of 15 ms in pre-boundary lengthening and 45 ms of pause (for a total of 60 ms per interval, or 360 ms maximum lengthening). These results showed a strongly significant effect of the Second boundary ($F = 25.572$, $p < 0.001$), while the First boundary ($F = 1.364$, $p = 0.791$) and the interaction of factors ($F = 1.364$, $p = 0.90$) were both statistically not significant. MANOVA tests show strong significance ($p < 0.001$) for the Second boundary within levels of the First (except for First = 6), but no effect of First within the Second boundary.

The Mean Response score for token (0,0) was of -2.33, suggesting the presence of strong Overall Bias in favor of the early break interpretation. However, the Overall Bias effect is not visible in the Contour Line of the Distribution Plot shown below in Figure 45, as the strong Second Boundary Conditional Bias effect¹⁷ overshadows the Overall Bias (which would move the diagonal Contour line towards the top right corner, thus extending the early break area below the line).

¹⁷ The pattern of the Contour Line in Figure 45 is interesting, as at first sight it appears to violate our predictions about the interaction of Overall and Conditional Bias, and could not be explained by the processing hypothesis we put forth in this thesis. We propose instead that the diagonal portion of the Contour Line (below First = 4) is not a reflection of the Overall Bias, but rather an artifact created when trying to overlay a slightly slanted line (Second boundary C.B.) on a grid composed of strictly perpendicular lines. It would be necessary to extend the grid downwards, creating items with larger First Boundary sizes and the same Second boundary sizes, to confirm this explanation.

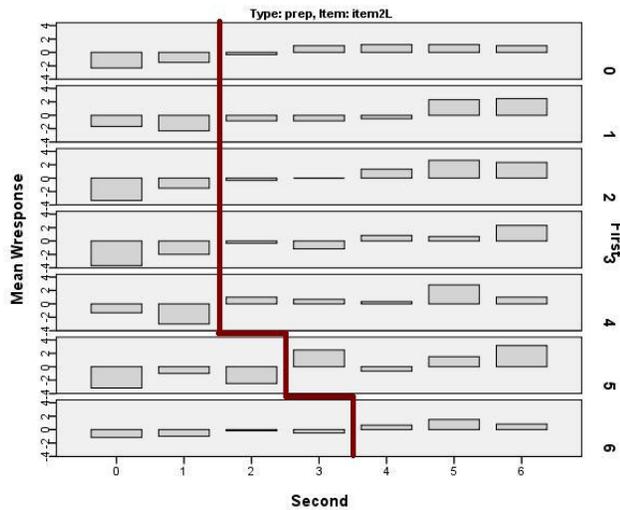


Figure 45: Distribution Plot for Item "With the Teddy Bear" (long, 7)

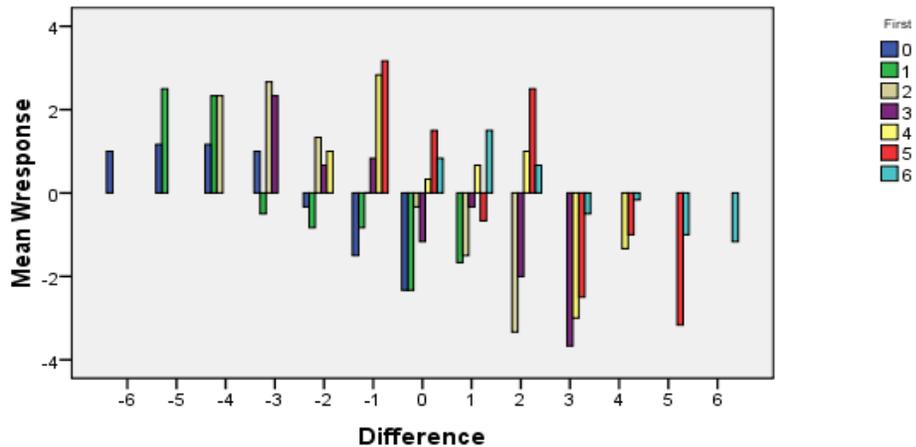


Figure 46: Difference Bar-chart for Item "With the Teddy Bear" (long, 7)

The Difference variable is also strongly significant ($p < 0.001$, $F = 6.148$), and MANOVAs show a statistically significant effect of the First/Second boundary within most levels of the Difference factor, with the exception of -6, -2, 2, 4 and 6. This can be confirmed graphically through the systematic variability within the cluster levels -2 to 2, while items outside this overlap zone are internally consistent with respect to interpretation choice (positive or negative score).

Given the strength of these results, we decided to keep the same phonetic properties of this larger-range grid (max 360 ms lengthening instead of 180) but to reduce the number of steps, so that the matrix would consist of fewer tokens (16 instead of 49), which would allow us to run the same number of subjects on a larger number of different items.

To confirm our results and facilitate comparisons across items, this item (*with the teddy bear*) was also re-run on the new 4-by-4 grid. The results are consistent with those just reported for the same phonetic properties but finer grained intervals: the Second boundary is still strongly significant ($F = 10.653$, $p < 0.001$), and both the First boundary factor ($F = 0.692$ and $p = 0.559$) and the interaction of factors ($F = 0.442$, $p = 0.909$) are not significant.

The Mean Wresponse score for this token (0,0) is -1.5, again suggesting the presence of strong early break Overall Bias in the data, but given the interaction of this strong Second Boundary Conditional Bias, triggered by boundaries of 120 ms or longer, the graphical representation of the Overall Bias is obscured.

The Difference variable is significant ($p = 0.012$, $F = 2.854$), and MANOVAs show no statistically significant effects of the First/Second boundary factors, although the values approximate significance (set at $\alpha = 0.05$) for levels -6, 0 and 4. In Figure 48 below we can observe the systematic distribution of the Conditional Bias variation within clusters, as the first bar in Difference levels 0 through 6 patterns consistently differently from the rest.

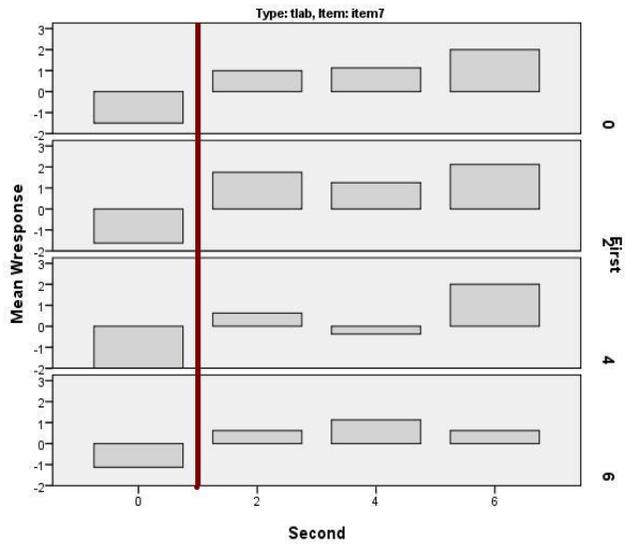


Figure 47: Distribution Plot for Item "With the Teddy Bear" (long, 4)

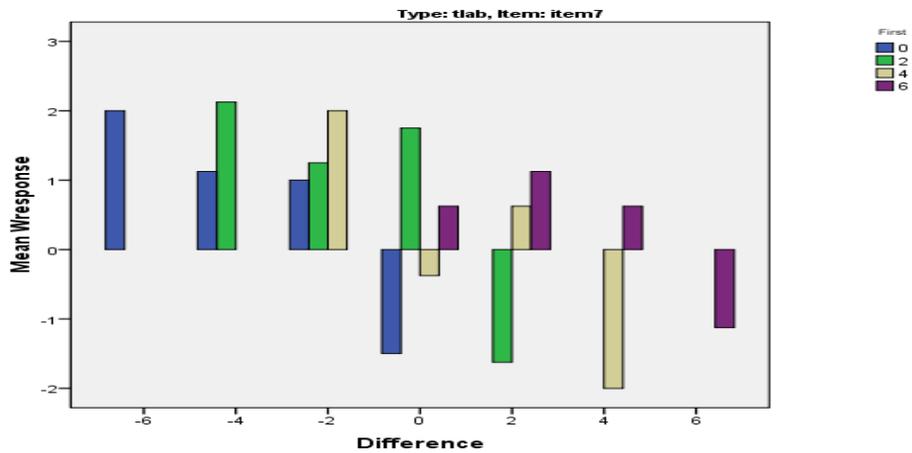


Figure 48: Difference Bar-chart for Item "With the Teddy Bear" (long, 4)

5.2.2 Rottweilers

Item (19b), and all those that follow, were only tested on the four-by-four long interval matrix that was just described. This item showed significance for the Second boundary factor ($F = 8.371$, $p < 0.001$), but not for the First ($F = 0.677$ and $p = 0.568$) or for factor interaction ($F = 0.456$, $p = 0.901$). MANOVAs show that the Second

boundary approaches significance within levels 0 and 2 of the First boundary ($p = 0.052$, $p = 0.032$ respectively) and nowhere else.

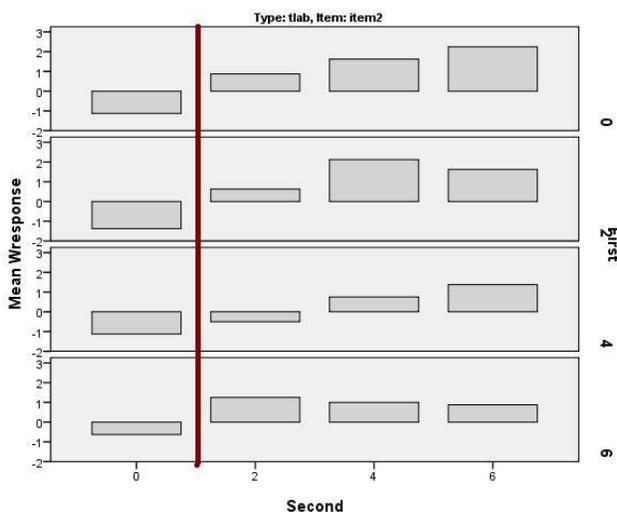


Figure 49: Distribution Plot for Item "With the Rottweiler"

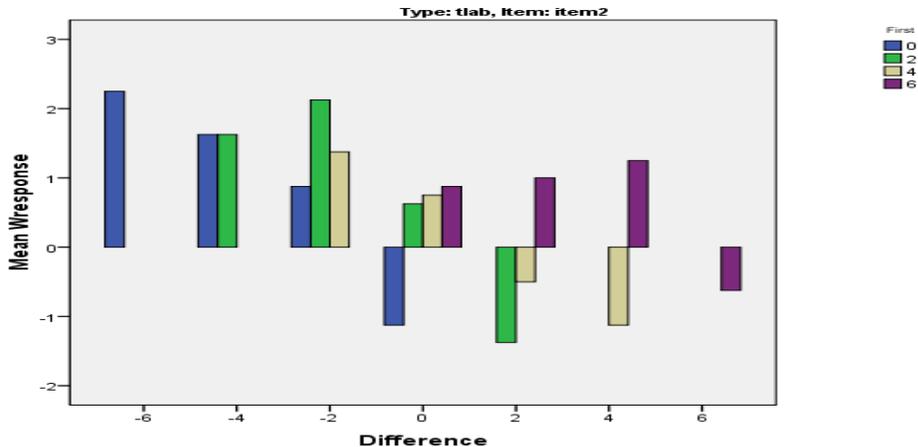


Figure 50: Difference Bar-chart for Item "With the Rottweiler"

The Mean Wresponse value for token (0,0) for this item is -1.125, which under our assumptions again suggests the presence of Strong Overall Bias favoring the early break interpretation. This result is however obscured by the strong Second boundary

Conditional Bias present, for all tokens with a Second boundary larger than 120 ms, as was the case for the previous item as well.

The Difference variable is also statistically significant ($p = 0.013$, $F = 2.835$), and while the MANOVA tests do not show any statistically significant effect of the First/Second boundaries within levels of the Difference variable, there is clearly variation in Difference levels 0, 1, and 2, which are the only items in which the Second boundary level 0 (responsible for most of the negative-score observations) is present.

5.2.3 Worried Expressions

This item again displayed the same results as the previous two: statistical significance for the Second boundary with $F = 3.127$ and $p = 0.029$, but not for the First boundary ($F = 2.504$ and $p = 0.063$). The interaction of factors was not significant ($F = 0.852$, $p = 0.570$), and there are no statistically significant effects of either boundary within the other, as reported by MANOVA tests.

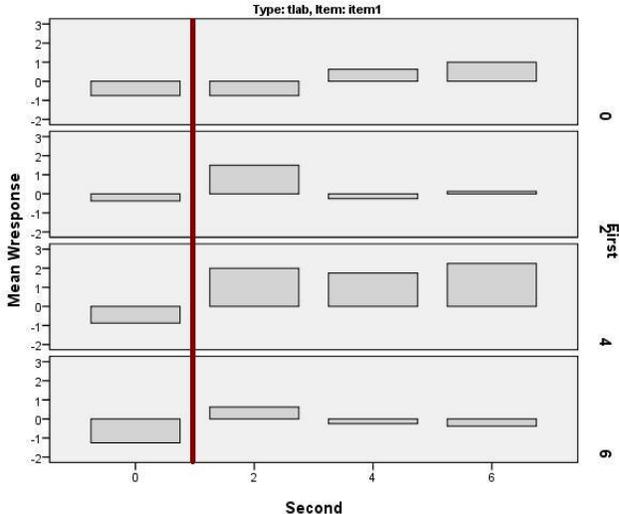


Figure 51: Distribution Plot for Item "With a Worried Expression"

The mean Wresponse score for token (0,0) is of -0.75, which under our assumptions does not qualify as suggesting the presence of strong Overall Bias for this item. Graphically, it is impossible to verify either way, since the extremely strong Second boundary Conditional Bias creates a perfectly vertical contour line separating tokens with a Second boundary of 120 ms or larger from those smaller in size.

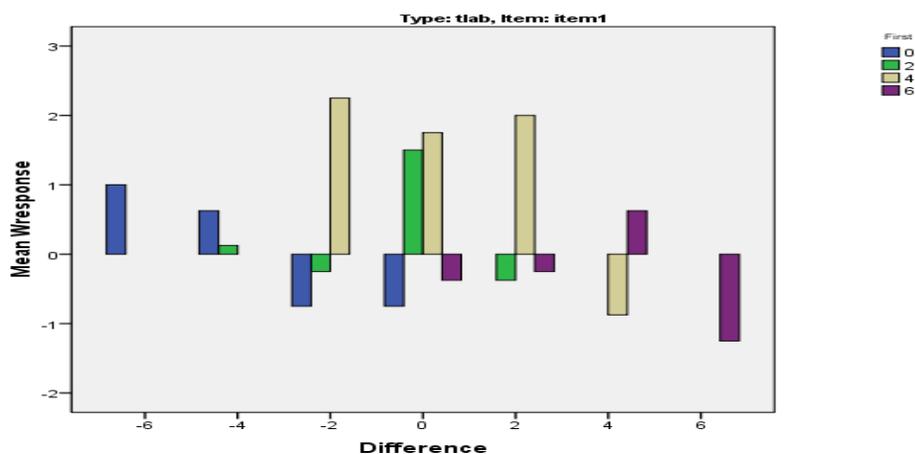


Figure 52: Difference Bar-chart for the Item "With a Worried Expression"

The Difference variable is also not significant ($p = 0.625$, $F = 0.731$), and Figure 52 shows that this might be due to the large amount of variability within and across the levels of the Difference variable. Despite the noise confusing the results, it is still possible to detect systematic differences between the first bar of the clusters 0 through 6, and the other items. The distribution is clearer in the contour plot in Figure 51, but note that there are some exceptions to the distribution.

5.2.4 Offending Bows

This and the following two items behave slightly differently from what has been discussed so far. Although the statistical results are similar, a factor which may

be driven by the low number of levels within each factor and hence smaller number of subgroups to compare, the graphs displaying the distribution of the Wresponse variable tell a different story.

Item (19d) displays statistical significance with respect to the Second boundary factor ($F = 9.483$ and $p < 0.001$), but not with respect to First ($F = 0.845$, $p = 0.472$) or for the interaction of factors ($F = 1.457$, $p = 0.173$). MANOVA tests show effects of the First boundary within the Second at levels 0 and 6 ($p < 0.02$), and of the Second boundary within the First at levels 2, 4, and 6 ($p < 0.05$).

The Mean Wresponse score for token (0,0) of this item is of 0.5, which does not qualify under our assumptions as suggesting the presence of strong Overall Bias for this item, as is supported by the position of the Contour Line in Figure 53.

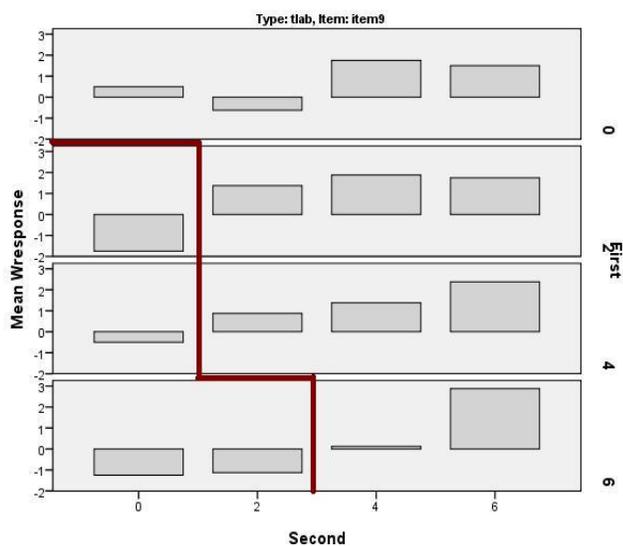


Figure 53: Distribution Plot for Item "With a Bow"

The Difference variable is significant at $p = 0.001$ ($F = 4.045$), and MANOVA tests show that there is significant variability attributable to the First/Second boundaries only at Difference levels -2 and 6.

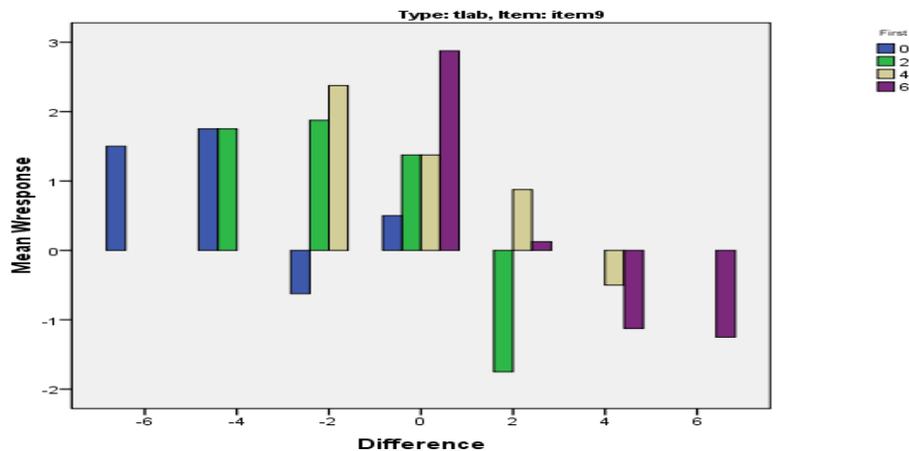


Figure 54: Difference Bar-chart for Item "With a Bow"

However, in looking at Figure 54, it is clear that there is no systematic variation between items, and that the Difference levels from -6 to 0, and 4 to 6, pattern separately, suggesting that there is no effect of Conditional Bias affecting this item. At the same time, the low Mean Wresponse score for token (0,0), as well as the centered Contour Line in Figure 53, which is diagonal and borders the edge token (0,0) suggests that there is little or no effect of Overall Bias either.

This item could therefore be said to display a distribution of responses that reflects almost exclusively only the relative sizes of the two boundaries in question. The very slight preference for late break interpretations, represented by the Contour Line's slight slant towards the bottom left corner of the Distribution Plot in Figure 53, could be attributed to either very slight effect of either Overall or Conditional Bias, but it is not possible to determine which at this stage.

5.2.5 Cannons

This item behaves in a similar fashion, although it actually shows statistical significance for both the First and Second boundary factors ($F = 3.990$, $p = 0.011$; and $F = 4.496$, $p = 0.005$ respectively). There is no interaction of factors, with $F = 1.008$

and $p = 0.438$. MANOVA tests show significant effects of First within Second at level 0 only, and of Second within levels 4 and 6.

The Mean Wresponse score for token (0,0) is of approximately 0, although as can be clearly noted from Figure 55, it is artificially low due to noise or perhaps problems with the specific token, since all three neighboring tokens show extremely high values for the Mean Wresponse score, which would be indicative of a strong Overall Bias in favor of late break interpretations, which can also be observed in the Contour Line of the Distribution Plot.

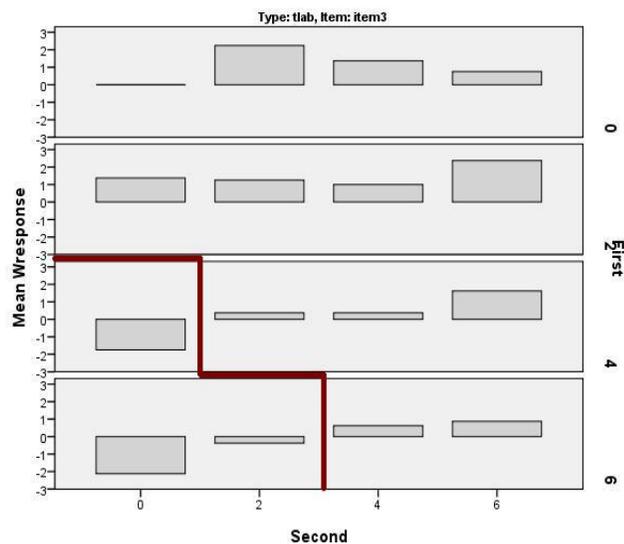


Figure 55: Distribution Plot for Item "With the Cannon"

The Difference variable is significant at $p < 0.001$, with no statistically significant variation (except for levels 4 and 6), and in Figure 56 it is possible to see that all variation is clearly contained within the same response type (i.e. it is variation in the confidence of the score, rather than in the actual break decision), and both Figure 55 and 56 show that this is a very clear case of Overall Bias shifting the point at which boundaries are considered to be perceptually equal (that is, when the meaning

crosses over from early to late break interpretation), to between Difference levels 2 and 3.

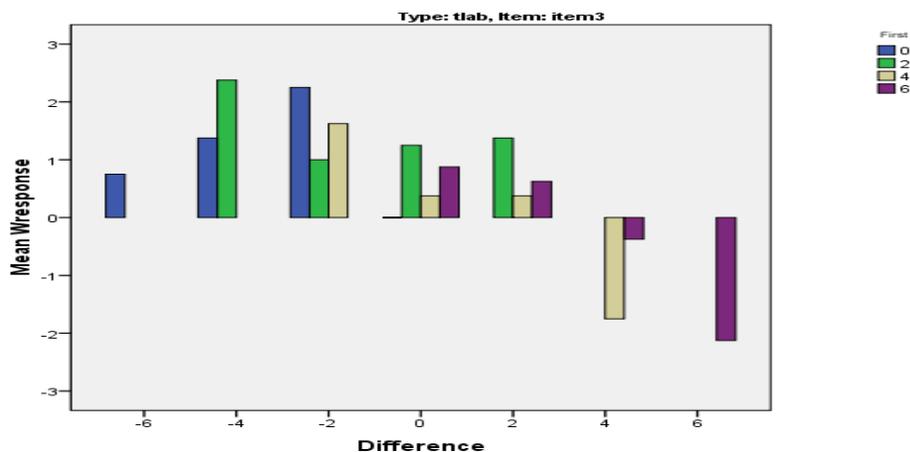


Figure 56: Difference Bar-chart for Item "With the Cannon"

Due to the presence of this strong Overall Bias towards late break interpretations, the Second boundary Conditional Bias effect (which also would favor late break interpretations) is masked and it is impossible to determine whether it is applying redundantly for this item or not.

5.2.6 Attack Plans

This item behaves similarly, showing strong significance of the Second boundary ($F = 4.873$, $p = 0.003$), while the First boundary approaches significance ($F = 2.213$, $p = 0.091$), but there is also a strong interaction of factors ($F = 2.824$, $p = 0.005$). MANOVA tests show statistically significant effects of the First boundary within levels of the Second (with the exception of Second level 4), as well as of the Second boundary within all levels of the First with the exception of First level 6.

The Mean Wresponse score for token (0,0) for this item was of -0.75, which under our assumptions does not suggest the presence of strong Overall Bias in either direction.

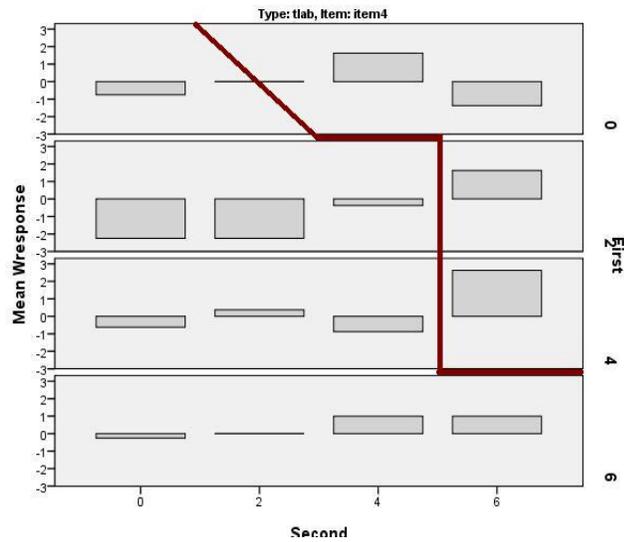


Figure 57: Distribution Plot for Item "With the Attack Plan"

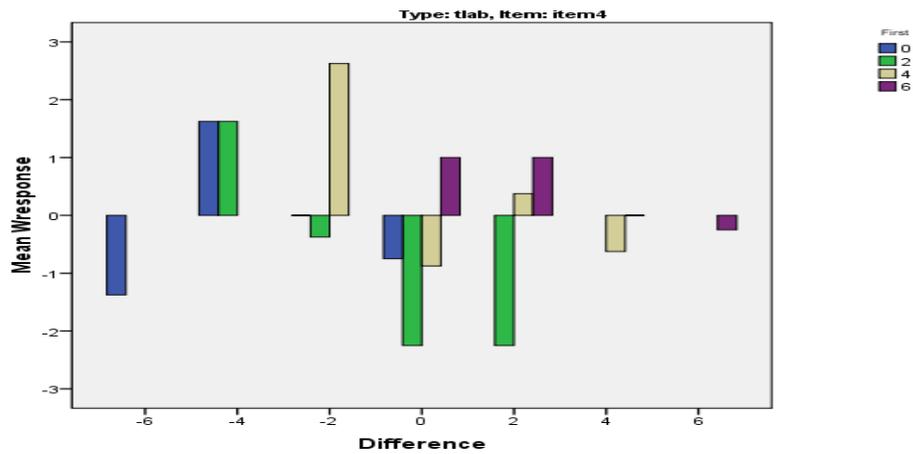


Figure 58: Difference Bar-chart for Item "With the Attack Plan"

The Difference variable is significant ($p = 0.016$, $F = 2.724$), and MANOVA tests show statistically significant effects for Difference levels -6, -4, -2 and 0. While

Figure 57 suggests that the distribution of Wresponse data follows a simple shift of the contour line consistent with Overall Bias favoring a late break interpretation, the amount of noise contained within the clusters in Figure 58 suggests that the contour distribution in the first graph may be too optimistic.

5.3 Discussion

As predicted by our processing model, the items presented in this study displayed results consistent with a Second boundary Conditional Bias results distribution. For all items displaying Conditional Bias effects, tokens with Second boundaries of 120 ms or larger were consistently assigned late break interpretations. Furthermore, these results are consistent with the results extracted from the first item, run on finer-grained (60 ms windows instead of 120), and in that case (discussed in 5.2.1) the contour line occurs between levels 1 and 2, corresponding to the window between 60 and 120 ms total duration of boundary lengthening. The consistency across trials of the *Teddy Bear* item, as well as across other prepositional attachment structures in which it is observed, suggests that the phonetic properties of a boundary triggering Conditional Bias may be generalizable across structures, if not even universal.

However, not all items did display Conditional Bias effects, which is problematic for the predictive power of our model, which would expect all items with the same structure and task to behave in a similar fashion. Two items, *Bow* and *Cannon*, displayed Overall Bias favoring the late break interpretation, and no effect of Conditional Bias, but this could be explained by redundancy, since Second boundaries of 120 ms or larger are for both these items already received late break interpretations from the Overall Bias.

Unfortunately, one problematic item remains: the *Attack plan* item, discussed in section 5.2.6. The distribution of results suggests an Overall Bias towards the early break interpretation, and which clearly does not show any effect of Second-boundary related Conditional Bias. Unlike the previous two items, this means that a number of tokens with a Second boundary 120 ms or longer (in some cases much longer) were receiving early break interpretations, contrary to our hypothesis. We saw that in previous sections as well there were a few items that did not display any effects of the Conditional Bias that we predicted and were realized for sister items, and further research is necessary to determine whether these represent a problem with the hypothesis, the experimental method, or the stimuli selection and creation.

STUDY SIX: RELATIVE CLAUSES

6.0 Introduction

We next consider a structure that was also traditionally considered a strong example of syntactic ambiguity that could be resolved by prosodic information, and which was extensively tested by Clifton, Carlson and Frazier (2002): Relative Clauses. In these structures, the relative clause *wh*-phrase could be interpreted as attaching to either the higher or lower noun of a possessive structure (*the X of the Y*).

Given the revised nature of the task to facilitate processing of the sentence's ambiguity and meaning, we predicted a First boundary Conditional Bias effect throughout the data, which was strongly realized on both items tested.

6.1 Method

Informal observations of these items showed that these too, like the *with*-phrases of the previous chapter, tend to be heavily influenced by lexical and contextual bias. Furthermore, the items are longer than any tested so far, which appeared to affect the ease of processing by the participants, so the experimental task was slightly modified to accommodate that.

6.1.1 Stimuli

The Relative Clause stimuli were taken directly from the Clifton, Carlson and Frazier (2001) paper, where they had been shown to be sensitive to prosodic information for meaning resolution:

27. a. I met the daughter | of the colonel | who was standing on the balcony
- b. Pam saw the killer | of the journalist | who got a lot of media attention

6.1.2 Subjects

Participants were recruited from the Cornell undergraduate population and were required to speak American English as their native language. They were compensated with \$5 or an extracredit point in their undergraduate psychology class for their time. Six subjects were run on the first item, and four on the second.

6.1.3 Experimental Setup

As before, participants were asked to select both a possible response to the target question, which would disambiguate the sentence, as well as confidence rating, which were used to calculate the Wresponse score results. Given the length and complexity of the structure of the relative clause sentences, it was deemed unnecessary and excessive to provide subjects with a further continuation of the sentence, and we decided that it would be easiest to paraphrase the relevant section of the ambiguity:

28. I met the daughter of the colonel who was standing on the balcony.
 - a. The colonel was standing on the balcony.
 - b. The daughter was standing on the balcony
29. Pam saw the killer of the journalist who got a lot of media attention.
 - a. The journalist received lots of media attention.
 - b. The killer received lots of media attention.

Subjects were familiarized with both possible interpretations in a written training session before the start of the experiment, and they were asked to report any oddness or ungrammaticality to the experimenter.

6.1.4 Conditional Bias Predictions

Changing the nature of the task, while beneficial in lowering processing time and participant frustration, also altered the way in which the sentence was parsed by

listeners. Rather than picking the most appropriate completion based on the global sentence meaning, speakers were made aware of the ambiguity and were asked which noun the relative clause modified—effectively changing this task into a pairing-task similar to that seen for simple and modified conjunctions in Studies Two and Three.

For this reason, we would predict that speakers could already form an opinion about the relative grouping of the Noun-Noun-RC constituents as early as the first boundary location, which occurs right after the first noun in the possessive chain. A First boundary Conditional Bias would predict that for any First boundary larger than a certain value, the probability of a late break interpretation would be significantly reduced.

However, even if the task had remained one of sentence continuation or if we had used another non-pairing method, we would still predict a First boundary Conditional Bias effect, since the section of the sentence ending at the first boundary location (*Pam saw the killer*) can be processed at that point for meaning which would affect the attachment location of the final relative clause phrase.

6.2 Results

As before, the Wresponse distribution was analyzed using ANOVA and MANOVA statistical tools to determine the effect, if any, of the First, Second and Difference boundary factors. Significance was set at $\alpha = 0.05$, although results that approach significance are also reported.

6.2.1 The Daughter of the Colonel

This item showed a strongly significant first boundary effect ($F = 3.822$ and $p = 0.001$), but no significance for either the Second boundary factor ($F = 1.019$ and $p =$

0.415), or for factor interaction ($F = 0.672$ and $p = 0.918$). MANOVAs showed no significant effect of either boundary within the other at any level.

The Mean Wresponse score for this item is 1.33, suggesting the presence of strong Overall Bias favoring a late break interpretation. However, given the low scores of the three neighboring items, it seems more likely that this score is attributable to noise or priming effects within the experiment structure than to true bias in the source file.

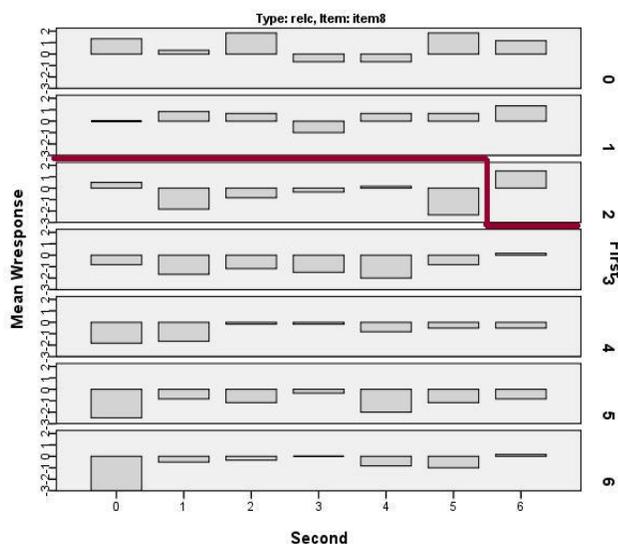


Figure 59: Distribution Plot for Item "Daughter of the Colonel"

The Difference variable is statistically significant, at $p = 0.017$, ($F = 2.105$). A MANOVA test shows statistically significant difference only at levels -5 and 6 of Difference, by the First/Second variable combination.

The distribution of the Wresponse data in Figure 59 clearly shows the effects of a strong First boundary Conditional bias¹⁸, for which virtually all First boundaries

¹⁸ As before, we believe that the slight dip at the end of the line is not due to a late-effect Overall Bias, but rather to the geometric properties of trying to overlay a straight but slanted line representing the First boundary Conditional Bias on a grid composed solely of

of size 2 (60 ms) or larger would significantly reduce the probability of a late break interpretation. Note, however, that even some tokens above the contour line often display a preference for an early break interpretation—this can be clearly seen in the systematic variation between the first two columns and the remaining columns in the Difference bar chart shown in Figure 60.

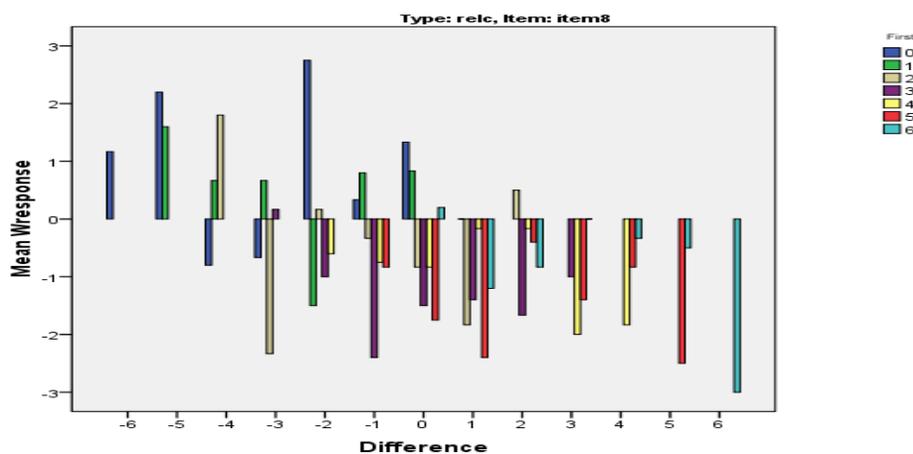


Figure 60: Difference Bar-chart for Item "Daughter of the Colonel"

6.2.2 *The Killer of the Journalist*

The same results are observed for item (5b), which also shows a highly significant effect of the First boundary factor ($F = 3.848$, $p = 0.001$) but not of the Second boundary ($F = 1.380$, $p = 0.223$), or of factor interaction ($F = 0.719$, $p = 0.883$). MANOVA tests show no statistical significance at any level of either boundary within the other.

The Mean Wresponse score for token (0,0) of this item is -1.5, suggesting the presence of strong Overall Bias towards an early break interpretation (or a high attachment of *media attention* to *the killer*, which is contextually justified). However,

perpendicular lines. Extending the matrix further to the right, testing items with larger Second boundary values, would confirm or deny our explanation.

both the fact that neighboring tokens, such as (1,0), have much lower scores, and the bizarre distribution of the Contour Line in Figure 61 suggest that this may be an artifact of noise or experiment priming, rather than a true reflection of the Overall Bias in the Item¹⁹.

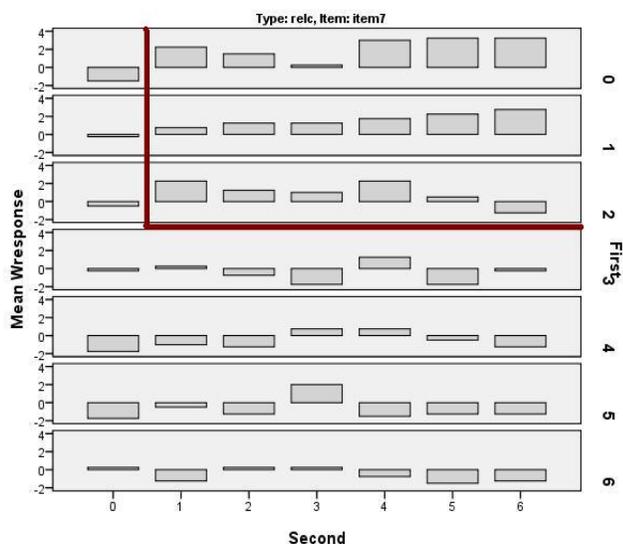


Figure 61: Distribution Plot for Item "Killer of the Journalist"

The Difference variable is barely significant at $p = 0.039$ ($F = 1.882$), and the First/Second boundaries are significant in affecting the Wresponse distribution within levels -6, -5, and -4 of the Difference variable.

As before, the distribution of the Wresponse variable within the First and Second boundary grid, shown in Figure 61, shows a very clear effect of First boundary Conditional Bias, with items with a First boundary of 90 ms or above displaying early break responses, even when the Second boundary is phonetically larger than the first. This effect reveals itself in the systematic variation in Wresponse values in the

¹⁹ Our theory of processing does not explain a right-angle Contour Line scenario as in Figure 61, and it would be necessary to run more tests with finer grained intervals and more subjects to confirm these findings and determine their causes, if any.

Difference plot in Figure 62, where the first three columns of each cluster pattern differently with respect to the remaining columns, for Difference levels -6 through 0.

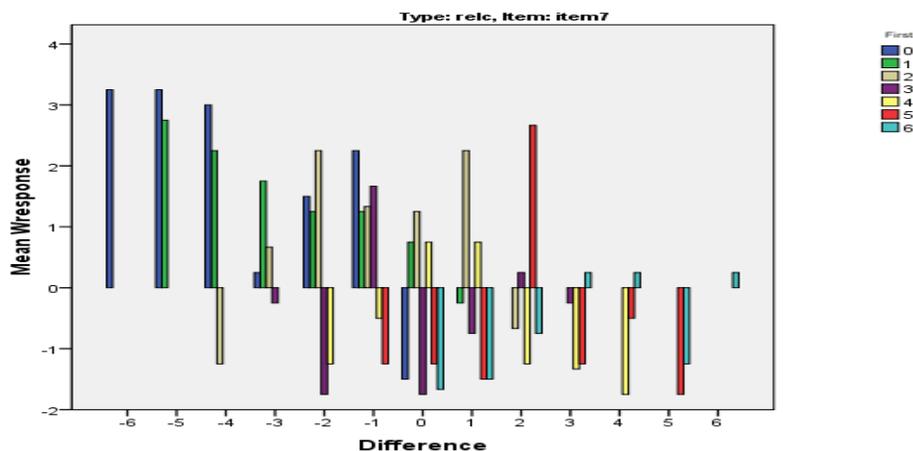


Figure 62: Difference Bar-chart for Item "Killer of the Journalist"

There are a number of other tokens in Figure 62 that display a late-break score, but as these do not appear to fit a logical pattern—note for example the alternation between early and late boundary meanings within level 2 of the Difference factor—it is difficult to describe them as anything but ‘noise’ in the data.

6.3 Discussion

Although only two items were tested for this structure, the consistencies across items are remarkable—both displayed extremely strong First boundary Conditional Bias, which appears to affect boundaries larger than 60 or 90 ms, and which significantly lowers the probability of a late break interpretation throughout the paradigm. These results are in line with the predictions of our model.

STUDY SEVEN: MIDDLE ATTACHMENTS

7.0 Introduction

Lastly, we test the strength of our processing model on a set of non-traditional attachment ambiguities, taken from Price et Al (1991). These items were dubbed Middle Attachment ambiguities, as the structure can be most easily visualized as having the middle constituent attaching to either the earlier or later constituent by left or right attachment. In some cases, in fact, the first and third phrases are not even part of the same sentence, and therefore could not be part of the same syntactic tree with the high/low attachment distinctions that have characterized the more traditional ambiguous structures.

These ambiguities present a number of interesting prosodic features, and yet they can still be disambiguated using virtually the same prosodic cues as were used for previous items, and display the effects of Conditional Bias effects as predicted by our model.

7.1 Method

For this experiment, we return to the sentence completion experiment paradigm that we had implemented in previous studies. Most other aspects remain unchanged.

7.1.1 Stimuli

As mentioned above, the middle attachment ambiguities studied were inspired by the items presented in Price et Al's (1991) paper. Items 30a and 30b were taken directly from the paper, but although additional items were considered, an informal survey of grammaticality judgments with non-naïve native speakers showed that for many items, one of the readings was considered ungrammatical or at least strongly

implausible. Item 30c was modeled after the items presented by Price et Al, and was judged to be acceptable by native speakers.

- 30. a. Although they were running | in the woods | they were uneasy.
- b. Since Tess will present | clearly | she will convince them.
- c. When you learn | gradually | you worry more.

All items were run on the standard 7 by 7 matrix with 30 ms steps between levels, divided equally between pause and duration cues.

7.1.2 Subjects

Participants were recruited from the Cornell undergraduate population and were required to speak American English as their native language. They were compensated for their time with \$5 or one extracredit point for their undergraduate psychology class. Six subjects were run on each of the items in this section.

7.1.3 Experimental Setup

As before, participants were asked to select both a possible response to the target question, which would disambiguate the sentence, as well as confidence rating, which were used to calculate the Wresponse variable results. The answer choices mirror those used for other items in this study, and consist of sentence continuations that would help disambiguate the intended scope of the adverb. The length of the target sentence and of the responses, especially when compared across tasks, may have been a factor in participants' processing of the sentence and interpretation choices.

- 31. Although they were running in the woods they were uneasy...
- a. ...but as soon as they got back into town, they calmed down
- b. ...which is strange, because normally running in the woods is relaxing

32. Since Tess will present clearly she will convince them...
- a. ...She's been taking oratory classes and is now a great speaker
 - b. ...if they had invited Bob instead, I'm not so sure he'd convince them
33. When you learn gradually you worry more...
- a. ... but if you learn quickly, then you're not concerned about it.
 - b. ... and you slowly realize how dangerous things can be.

Subjects were familiarized with both possible interpretations in a written training session before the start of the experiment, and they were asked to report any oddness or ungrammaticality to the experimenter.

7.1.4 Conditional Bias Predictions

Our processing hypothesis would predict a First boundary Conditional Bias for these items. The first boundary break location does not occur at what could be the utterance ending, as was the case for some earlier items (i.e. *The Vikings won*), but they are nonetheless possible constituent and phrase final points (i.e. *Since Tess will present*) at which processing of the preceding material can occur. More material is expected because of the sentence-initial subordinate conjunction (*Although, Since, When*), but this could very well be in the form of a completely disjoint phrase.

There have also been concerns raised that due to the length of the sentence and of the continuations, and the transparent attachment of the middle parenthetical or adverb to either phrase or sentence, speakers would have reinterpreted this task as a forced pairing situation. This situation, which we already examined in simple and modified conjunctions discussed in Study Two and Three, also predicts a strong First boundary Conditional Bias, as at the first boundary location listeners can already form an opinion about the relative pairing of the constructions, and given a large enough

boundary are likely to start building the early-break structure, thereby lowering the probability of a late-break structure at those points.

7.1.5 Peculiarities of the Structure

There is one further source of early break bias in this structure, created by the very particular prosody of the items selected to represent Middle Attachment structures. Consider the sentences:

34. a. Clearly, she'll convince them.

b. In the woods, they were uneasy.

In pronouncing these items, one can easily place a large pause or prosodic break at the comma location, after the initial adverb or phrase, and it would sound quite natural. If these sentences were preceded by further material, we would expect that the large prosodic break after the topicalized element (Second boundary location) would not necessarily cue attachment of the topicalized element to the preceding phrase.

However, we would predict that when larger than a certain size, prosodic boundaries in the post-topicalized position would be interpreted as true indicators of a break, and would be again understood as cues suggesting the presence of a late break.

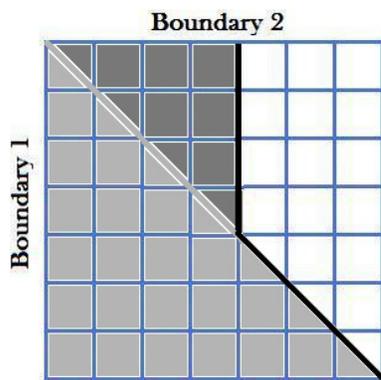


Figure 63: A Possible Graphical Representation of Topicalization-Bias

In the absence of Overall and Conditional Bias effects, we could represent this situation as shown in Figure 63, with the two shaded areas representing the early break decisions, divided from the late break area (white) by the dark contour line.

The normal distribution of early break decisions (below the diagonal, represented by the light grey shading) is compounded by the darker grey area, which represents cases in which the Second boundary is slightly larger than the first, but the second boundary is heard simply as a post-Topicalization pause, and is not entered into the calculation of relative boundary size and hence boundary location.

We expect Second boundaries larger than a certain size to resume their normal role indicating the end of the phrase, but we do not expect the size of the First boundary to have any effect on the Topicalization bias, and describe this graphically as the vertical Contour line the descends until the diagonal (shifted, if necessary, to indicate Overall Bias) and then continues. If present, the First boundary Bias (represented by a horizontal Contour Line) could extend to the right from either the Topicalization vertical line, or from the Overall Bias diagonal, depending on the relative sizes of the boundaries at which the Topicalization and Conditional Biases take effect.

7.2 Results

As before, we report on a number of statistical tests run to assess the influence of the First, Second, and Difference factors on the distribution of the Wresponse (weighted response) score. These include the standard Analysis of Variance (ANOVA), as well as the Multivariate Analysis of Variance (MANOVA), which performs a series of ANOVAs for one factor within levels of the other. The factors analyzed here (First, Second, Difference) are considered Fixed for statistical purposes unless otherwise stated; and significance was set at $\alpha = 0.05$ throughout.

7.2.1 In the woods

This item shows clear significance for the First boundary factor ($F = 4.784$, $p < 0.001$), and no significance for either the Second boundary ($F = 1.099$ and $p = 0.363$) or for the interaction of factors ($F = 0.719$, $p = 0.883$). MANOVAs show no statistical significance of any factor within any level of the other.

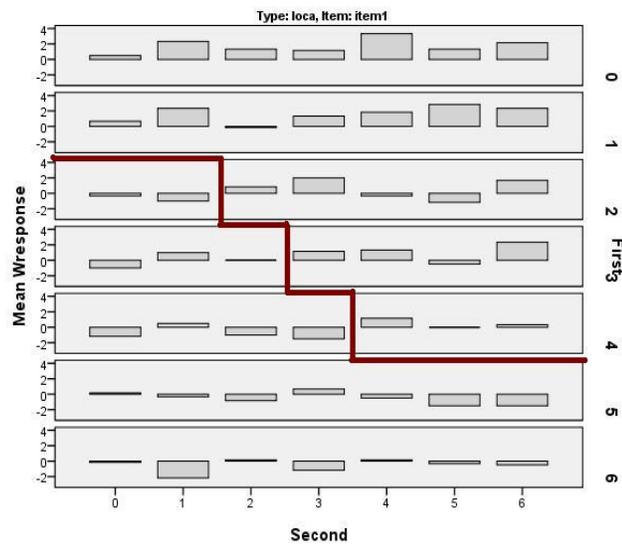


Figure 64: Distribution Plot for Item "In the Woods"

The distribution of the Wresponse variable in Figure 64, and in particular the contour line, suggest a First boundary Conditional Bias, consistent with our predictions, affecting tokens with a First boundary of 120 ms or longer.

The Mean Wresponse value for token (0,0) is 0.5, which under our assumptions does not qualify as representing a strong Overall Bias in either direction. This is supported by the diagonal portion of the Contour Line in Figure 64, which runs close to the center of the figure, where the Difference = 0 tokens lie.

The Difference variable is significant ($p = 0.004$, $F = 2.471$), and MANOVAs show that the First/Second boundary factors do not have any significant effect in the distribution of the Wresponse variable within levels of the Difference variable.

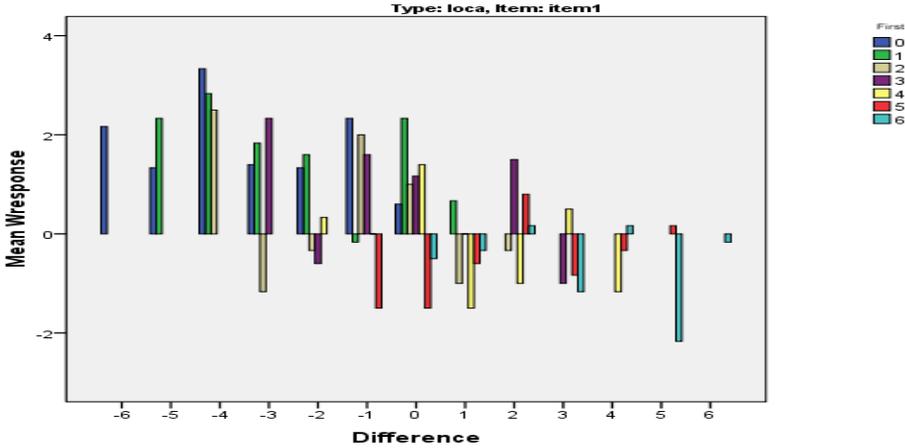


Figure 65: Difference Bar-chart for Item "In the Woods"

The distributions of responses within the Difference levels, shown in Figure 65, displays some noise in the more extreme levels (-3, 2, 3) but otherwise shows systematic variation within levels -2, -1, and 0, with more extreme positive and negative Difference factor levels displaying internally consistent early- and late-break interpretations, respectively.

7.2.2 Clearly

Item (6a), tested on six subjects, approximated significance in the distribution of the First boundary factor ($F = 1.852$, $p = 0.090$), with a clear absence of statistical significance for the second ($F = 1.125$ and $p = 0.348$), and interaction of factors ($F = 0.522$ and $p = 0.989$). Neither boundary has any effect on either level of the opposing variable.

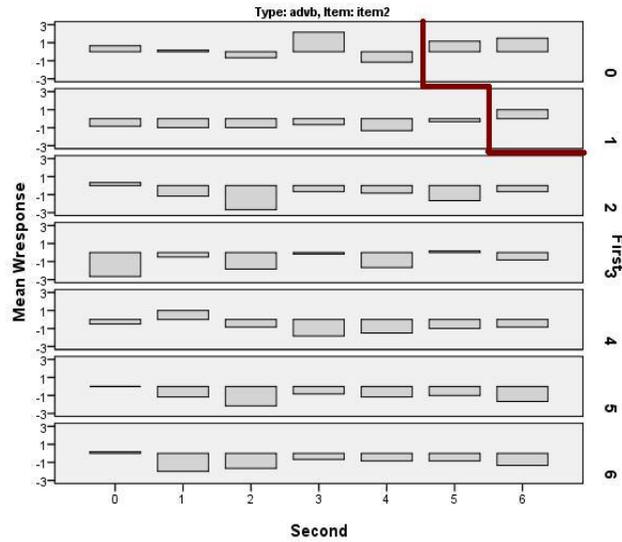


Figure 66: Distribution Plot for Item "Clearly"

The weakness of our crude assessment for Overall Bias is most apparent in this item. Here, the Mean Wresponse score for token (0,0) is a mere 0.66, which not only does not indicate the presence of strong Overall Bias, but is also slightly slanted in the wrong direction, indicating a late break interpretation in the midst of a very large area of early break decisions. It seems that overall, but especially in the top-left quadrant, subjects were particularly unsure about which item to select, choosing either low confidence judgments, or as in the case of the token (0,0), selecting both possible interpretations with equal frequency (the unweighted Response score was 0), but giving slightly higher confidence ratings to one or the other interpretations which ultimately affected the final Wresponse rating.

The Distribution Plot in Figure 66 however clearly illustrates that there is a strong Overall Bias in favor of early break interpretations, with the Contour Line shifted towards the top-right corner, for which we would have expected a large negative value for the Mean Wresponse score of token (0,0).

The Difference variable is also not statistically significant as an explanatory variable for the distribution of the Wresponse variable ($F = 1.100$, $p = 0.360$) and a MANOVA test shows no significant effect of the First/Second boundaries within any level of the Difference variable.

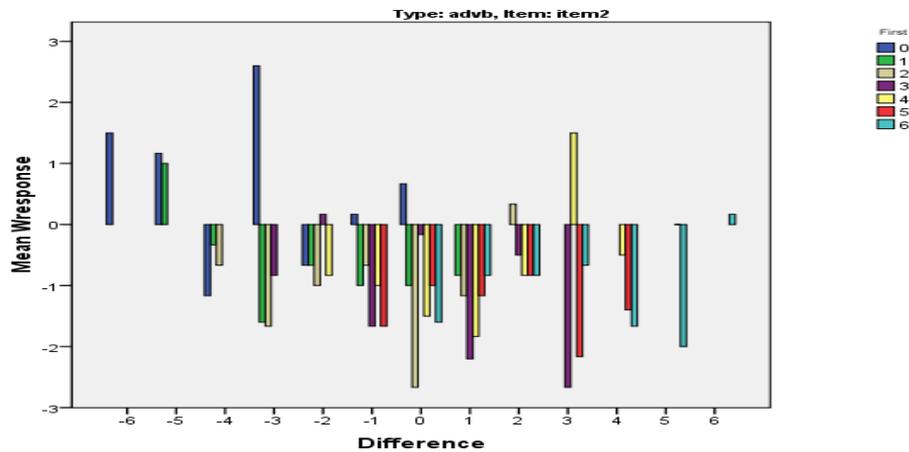


Figure 67: Difference Bar-chart for Item "clearly"

The distribution of responses within and across levels of the Difference factor, with the crossover point between interpretations pushed to the very end of the Difference variable (between levels -4 and -5), supports the strong Overall Bias interpretation of the data distribution. No Conditional Bias effects are detected, possibly because the items with a large enough First boundary to trigger the Conditional Bias are already judged to have early breaks due to the overwhelming Overall Bias affecting the data distribution.

The strong Overall Bias is probably also responsible for the lack of statistical significance throughout, as the highly prevalent early break interpretation means that all the scores would be concentrated between -4 and 0, reducing the possible variability between items.

7.2.3 Gradually

This item displayed a very interesting distribution of results which, although not particularly relevant to the model described in this paper, suggests avenues for further research and refining of the model. Neither the First, nor the Second, nor the Difference factors displayed any statistical significance ($p > 0.300$) for this item, due to the extremely limited number of late break responses (and the poor confidence scores these received).

What is interesting about this item, is that for a number of tokens with a stronger Second boundary than First, the item was still perceived as having an early break. The contour line in Figure 68 below suggests that at smaller levels of the Second boundary, the item displays a predictable distribution based on the relative size of the First and Second boundaries.

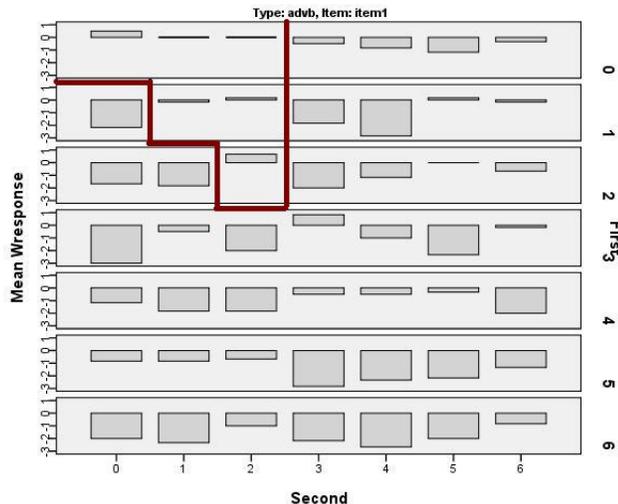


Figure 68: Distribution Plot for Item "Gradually"

However, for Second boundaries of size 3 or above (corresponding to about 90 ms), the probability of a late boundary attachment is significantly reduced. One possible explanation would be that boundaries above a certain size, in this context, are

reprocessed by the speaker as speech errors, and cancelled out from the processing, thus triggering an early break interpretation. Further research with more targeted items and structures would have to be conducted in order to test this hypothesis.

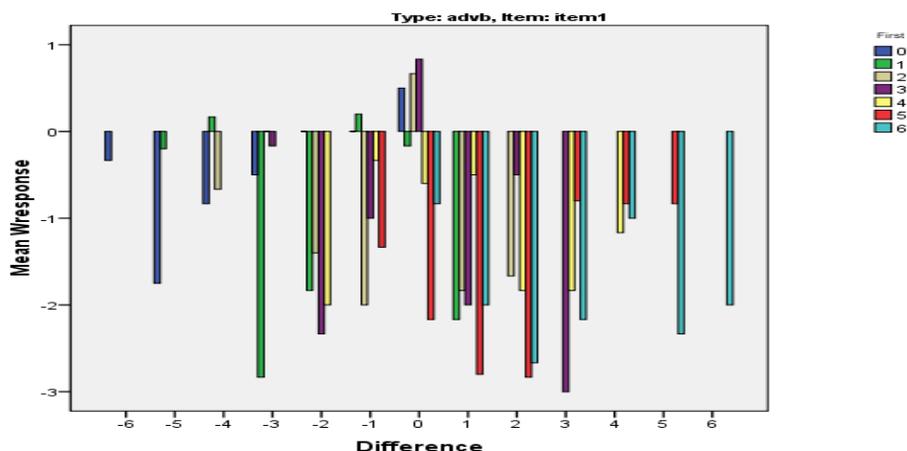


Figure 69: Difference Bar-chart for Item "Gradually"

It is impossible to determine whether the First boundary Bias would have applied in this structure, as the tokens it would have affected all received early boundary interpretations given the upswing of the Contour Line between Second levels 2 and 3.

7.3 Results

Interestingly, even though the items can be perceived as being part of two separate sentences, which may have reflexes on theories that relate prosodic disambiguation to the properties of the underlying syntactic structure²⁰, these items behave consistently with the model and Conditional Bias predictions presented at the beginning of the chapter.

20 Such as Clifton, Carlson and Frazier's Informative Boundary Hypothesis.

The first item, *in the woods*, presented clear First boundary Conditional Bias, as we had predicted, triggered by boundaries 120 ms or larger, while the item *clearly* had such strong Overall Bias towards an early break interpretation that the Conditional Bias effect appears to be redundant.

The final item tested, *gradually*, presented an interesting set of results which could be explained by the Second boundary being perceived as a speech error. However, since the same sized boundaries were tested on similar items without such an effect (and longer sized boundaries were successfully tested on other structures, as well) further research should be conducted to better map out the phonetic properties and constraints of this speech-error threshold.

Despite our predictions of a possible Topicalization bias, which would have turned a some of the tokens with a larger Second boundary into early break interpretations, we did not see any effects of this across any of the items analyzed in this section. It is possible that this option is over-ridden by a more classical interpretation of boundaries in an experimental context which explicitly tests the disambiguation of sentences.

GENERAL DISCUSSION

8.0 Summary of Results

The studies discussed in this thesis covered six different syntactic structures and a total of twenty-three different items. We include here a brief summary of the results before launching into a discussion of the findings and their implications for our model of prosodic processing.

Item Name	MWresp OB	Actual OB	Predicted CB	Actual CB
Formula B+C*D	No	No	2	2
B C D (short)	No	No	2	1
B C D (long)	No	No	2	1
Rose Steve (short)	No	?	2	1
Rose Steve (long)	No	No	2	1
Eve Jude Sue	No	No/slight	2	NO
Dancers Skaters	Yes (late)	NO	2	1 (weak)
Farmers Workers	No	YES	2	Redundant
Chefs Winetasters	No	YES (early)	2	Redundant
Check In (short)	No	No	1	NO
Check In (long)	Yes (late)	n/a	1	1
Drop Off	No	No	1	1
Win Over (short)	No	No	1	1
Win Over (long)	Yes (late)	n/a	1	1
Wear Down	Yes (early)	n/a	1	1
Look Up	Yes (early)	Yes (early)	1	n/a
Teddybear (short)	Yes (early)	n/a	2	2
Teddybear (long)	Yes (early)	n/a	2	2
Rottweiler	Yes (early)	n/a	2	2
Worried expressions	No	n/a	2	2
Bow	No	No/slight	2	Redundant
Cannonballs	No	Yes (late)	2	Redundant
Attack Plan	No	No	2	NO
Daughter of the Colonel	Yes (early)	n/a	1	1
Killer of the Journalist	Yes (late)	n/a	1	1
In the woods	No	No	1	1
Clearly	No	Yes (early)	1	Redundant
Gradually	No	No	1	Redundant

Figure 70: Summary of Overall and Conditional Bias Results

Figure 70 includes a summary of the predicted and actual results, for both Conditional (CB) and Overall Bias (OB), for all items tested. Incongruencies between the predicted and actual results are highlighted, and will be discussed in the following sections as we summarize the model, its predictions and limitations.

The first items run were simple tripartite conjunctions, in which three equally sized constituents were linked by short conjunctions. The items were all predicted to show Second Boundary bias based on our processing model, but only the Formula *B plus C times D* displayed it in the results.

We then hypothesized that certain types of Forced-pairing experimental tasks, where subjects had explicit access to the attachment structure, would show a First boundary Conditional Bias effect, which was visible for two of the remaining items, *B and C and D*, and *Rose and Steve and Kim*. These two items were tested over both short (180 ms max lengthening) and long (360 ms max lengthening) ranges: the results were replicated, although not strengthened as we would have predicted, a result which should be investigated further.

Lastly, we tested the mixed-conjunction *Eve or Jude and Sue*, which we would have expected to show Second boundary Conditional Bias in keeping with the processing structure, or at most First boundary Conditional Bias in keeping with the experimental task constraints, and yet it displayed a distribution of results consistent with slight Overall Bias, but no trace of either boundary Conditional Bias.

Next we decided to test modified conjunctions, in which the third constituent modifies either the second conjunct (low attachment) or the conjunct pair (high attachment); this was again framed as a pairing task, and we again expected subjects to be sensitive to a First boundary Conditional Bias. The results were again mixed: one

item, *Dancers and Skaters*, displayed a lukewarm First boundary Conditional Bias effect, but neither *Farmers and Workers*, nor *Chefs and Wine-tasters* displayed any form of Conditional Bias. However, the Overall Bias present in both sentences was such that the Conditional Bias could have applied redundantly over items that were already processed having an early break interpretation. More items (perhaps with Overall Bias leaning in the opposite direction) should be tested to verify our model's predictions for this structure.

We then decided to upgrade to entire sentences, and began by testing particle verbs, considered good examples of ambiguity by Price et Al but otherwise rarely discussed. Our model predicted a First boundary Conditional Effect, based on the processing properties of the structure itself, and the results confirmed it strongly. Three items (*drop off*, *wear down*, and *win over*) displayed extremely strong First boundary Conditional Bias effects throughout, which were consistent across both small and large phonetic interval tests; a fourth item, *look up*, displayed such strong Overall Bias that there were virtually no late-break tokens anywhere, and Conditional Bias effects would have been redundant.

One item, *check in*, was problematic in the short interval form as it only displayed weak Overall Bias and no Conditional Bias at all, and there is no possibility that the Conditional Bias was made redundant by the effects of the Overall Bias. However, running the same item over larger phonetic intervals did result in a strong First boundary Conditional Bias, so the shorter-range results should be taken with a grain of salt.

One startling finding of this set of data was that the score of the (0,0) token (acoustically identical across iterations of the same item) seems to depend on the

properties of the experimental setup, and its score was magnified when there were fewer items overall and with that interpretation.

Our next structure was the classic instrumental-or-modifier ambiguity of *with*-phrases, for which we decided to test six different structures over large phonetic intervals. Most items showed a strong Second boundary Conditional Bias, as predicted by our processing model, and these included *Teddy Bear*, *Rottweiler* and *Lecture*. Two more (*Bow* and *Cannonball*) showed late-break favoring Overall Bias effects such that a mild Conditional Bias would have been partially obscured.

One item, (*attack plan*) was quite clearly insensitive to Conditional Bias and presented Overall Bias slightly favoring the early break interpretation, such that there could not be any redundancy affecting the results. However, this item did present a certain amount of noise in the results, suggesting that this item might have been problematic for other reasons.

Our initial tests (with the item *teddy bear* only) showed that the standard short-interval manipulations resulted in a random distribution of results, which were however corrected when we lengthened them to the 360 ms range, suggesting that speakers might have a minimum size threshold which correlates with constituent length. Earlier items did not seem to be affected by this problem, which correlates with the size (both in length and complexity) of the three constituents as well as the overall clause.

The next study briefly considered Relative Clause ambiguities, in which the relative clause can be considered as modifying either the lower or higher element of a possessor phrase—a structure intensively tested by Clifton, Carlson and Frazier (2001). The task reverted again to a forced pairing structure, and because of this we

expected a First boundary Conditional Bias, which was strongly realized across both items (*The Killer of the Journalist*, and *The Daughter of the Colonel*).

Lastly, we turned to an interesting structure discussed in Price et Al (1991)—Middle Attachment Ambiguities, which were interesting since they are more intuitively described as a left or right attachment of the middle constituent (an adverbial phrase) to the flanking phrases, which were often unrelated sentences. For both processing and experimental task reasons, we predicted a First Boundary Conditional Bias effect, which was clearly realized in one of the three items tested (*In the woods*).

Of the remaining two, the item *clearly* displayed extremely strong Overall Bias which would have made the Conditional Bias redundant, and the other, *Gradually*, displayed very interesting behavior suggesting a boundary “maximum” had been reached, which opens up a new direction for further investigation.

In attempting to shed more light into the complex issue of prosodic resolution of syntactic ambiguities, these studies have uncovered a number of interesting findings, which in turn raise issues that would be well worth exploring in further work.

8.1 Overall Bias

In selecting the stimuli to be synthesized and used in these experiments we selected items that we believed would be as neutral as possible, and as such our only prediction and hope would have been to have items with very mild Overall Bias effects.

In some cases, however, there was a visible Overall Bias effect (graphically represented by a shift of the diagonal contour line in the Distribution Plot), which we attempted to quantify by comparing the weighted response scores of the (0,0) tokens

across items. However, our “predictions” based on the Wresponse score were occasionally incorrect, as relying on a single tokens’ score was highly susceptible to outlier judgments affecting the score and swaying our prediction.

We chose to use the (0,0) token as our baseline, since using other tokens with equal boundary sizes (1,1; 2,2; ... 6,6) would have added the risk of Conditional Bias interference, and these tokens could have corresponded to different phonetic manipulations according to the specific details of the experiment (7- or 4- level, 180 or 360 ms maximum lengthening, etc.).

In future work, we recommend the use of an independent measure of Overall Bias, derived from non-participants’ ratings of the base file (rather than the 0,0 token, which had undergone some manipulations), without exposure to other tokens of the same item. Given our assumption that Overall Bias is primarily influenced by lexical and phonetic information, it would be useful to extract the relative lexical frequencies of two disambiguated items, when possible—i.e. comparing the use of *drop off* meaning “delivery” vs. “to fall off (of something)”. It would also be useful to better control the immediate phonetic environment of the boundaries, in order to minimize the phonetic/acoustic sources of Overall Bias. World knowledge, the other component of Overall Bias, is harder to quantify, and for this it might be best to rely on intuitive judgments of subjects taken from the same pool as the experiment participants.

8.2 Conditional Bias

Our model included very specific predictions about the location of Conditional Bias which were met by the majority of the items tested. After considering the effects of the experimental task on processing, we revised the model slightly for Forced-pairing type tasks, and our predictions at that point were met, or at least not contradicted, by all items tested.

No items with the same structure and experimental task displayed opposite Conditional Bias effects, but occasionally there was variability among items as to whether or not Conditional Bias was displayed. In several cases, the distribution of scores for an item reflected only the phonetic difference between boundary levels, with slight variations due to the effect of Overall Bias.

8.2.1 The Predictions of Our Processing Model

We originally selected these structures in order to test the relationship between prosodic boundary effects and syntactic structures, and expected to find variation across structures that correlated to syntactic properties, such as the depth or height of an attachment point, constituent sizes, head-complement relationships, etc.

The results do show variations across syntactic structures, but it seems that the experimental task structure can override syntactic structure tendencies and make listeners favor one boundary over the other—this is particularly clear in the contrast between the Conditional Bias results of items like *B plus C and D* (Second boundary) and *B and C and D* (forced-pairing task, First boundary).

The Conditional Bias therefore groups the items into three distinct categories: those with pairing-induced First boundary Conditional Bias, which includes some Simple and Modified conjunctions, and Relative Clauses; those with structure-induced First boundary Conditional Bias, which includes Particle Verbs and Adverbs; and lastly those displaying Second boundary Conditional Bias, which includes the remaining Simple Conjunctions, and Prepositional Phrases.

It goes without mentioning that the six syntactic structures examined in this thesis form only a subgroup of all the possible types of ambiguities available to be studied, especially now that given the results of the Adverbial items in Study Seven, Prosodic disambiguation might not have to be restricted to the classical high-vs-low-

attachment syntactically conditioned ambiguities discussed in classic papers. Future research should explore other structures considered possible sources of ambiguity, as well as their interaction with different experimental tasks that may alter the normal processing sequence of the sentence.

8.2.2 Phonetic Properties of Conditional Bias

Previous studies of Prosodic boundaries often tested extremely large boundary sizes, often over 500 ms, and with very large gaps between boundary levels when these were compared. Some studies, such as O'Malley 1973 did not consider pauses under 300 ms to be indicative of a boundary, and so having just shown the consistent presence of an effect with boundaries less than 180 ms (90 ms pause, 90 ms lengthening) in duration is in and of itself an important finding.

However, one advantage of testing such a finely-grained grid of boundary intervals is the ability to draw conclusions about the specific phonetic properties of the points at which participants' interpretations switch over. We extracted the values at which the contour lines plateau for each item displaying Conditional Bias effects, and considering the amount of variation in the input (structure, constituent length, etc.) found remarkable similarities in the crossover locations.

The Second boundary Conditional Bias items display the tidiest results, with the formula *B plus C times D* switching between 60 and 90 ms total boundary manipulations, while the Prepositional Phrases over the long intervals—across all five useful items tested—crossed over between 0 and 120 ms total boundary lengthening, while over the short intervals the crossover occurred between 30 and 60 ms. This suggests that at around 60-90 ms a boundary (in the Second location of a sentence) is considered long enough for processing to occur that would skew the probabilities of interpretations in a manner consistent with Second boundary Conditional Bias.

The First boundary Conditional Bias items can be artificially divided into two groups: those displaying bias as a result of the experimental task, and those displaying it as a result of their own internal structure. The former include the simple conjunctions like *B and C and D*, whose crossover points were messy across repeated trials, and showed up between 90 and 120 ms total boundary lengthening in one case, and 180-240 ms in the other. Also Modified Conjunctions, of which only the item *Dancers and Skaters* showed clear effects of Conditional Bias between 120-150 ms boundary sizes, and finally Relative Clauses, whose crossover point occurred between 30-60 ms for one item, and 60-90 for the other. All in all, this suggests that the point at which First boundaries are large enough to trigger Conditional Bias (for paired-structure tasks) is somewhere in the vicinity of 90-120 ms, although there is space for variation that should be investigated.

The remaining items for First boundary Conditional bias showed similar results: Particle verbs had crossover points between 120-150 ms, and between 0 and 120 ms when tested over long intervals for the item *Drop Off*, but between 30 and 60 ms for the items *Win over* and *Wear Down*. The Adverb *In the Woods*, the only one to show relevant results, also had a crossover point between 120 and 150 ms, suggesting that the two extremely short items are somehow “exceptional”, and that the crossover point for this set of items is also in the vicinity of 120 ms.

These findings suggest that, regardless of the source, First boundary Conditional Bias is triggered by boundaries of about 120 ms, while Second boundary Conditional Bias seems to be triggered by slightly smaller boundaries, of about 90 ms. Of course these are only impressionistic observations, but they do suggest that further research should be conducted in determining the asymmetries between Prosodic boundaries at different locations within the sentence, and their effect on ambiguity resolution and sentence processing.

While this thesis focused primarily on determining the minimum amount of boundary size necessary for an interpretation to change, another interesting line of research (prompted by the results of the item *Gradually* in Study Six) is the study of whether boundary sizes are also constrained by ceilings. At these ceiling levels, pauses or lengthening would no longer be interpreted as actual boundaries, but rather as pauses or hesitations, and would be discounted by the speaker during the processing sequence.

However, it is important to keep in mind that the results obtained here were all based on synthetic speech deprived of any pitch modulations, and requires that we proceed with caution when extending these results to the processing of real speech.

8.2.3 Variability in the Domain of Conditional Bias

Three of the items tested in this thesis did not display any effects of Conditional Bias at all. These were: *Eve or Jude and Sue*, *The tourist checked in the bags* (short), and *The soldiers tried to locate the rebels with the attack plan*; they all have different structures, different experimental tasks and different predicted Conditional Biases, while other items with the same structure and experimental setup showed clear effects of Conditional Bias as predicted.

Two possible explanations exist: one is that the phonetic intervals tested were simply not large enough to pick up on any possible Conditional Bias, but this seems unlikely given the strength and consistency of the phonetic properties of Conditional Bias across items and structures, as described in section 8.2.2.

The other option is that for these particular items, Conditional Bias simply did not apply. It is interesting to note that these items were also completely immune from Overall Bias (both as a MWresponse (0,0) score prediction, as well as an actual

observable effect): in other words, the score distribution of these items only and exclusively reflected the relative size of the two manipulated boundaries.

It seems likely therefore that in processing these items, subjects were excessively cautious and did not process them as actual sentences, but rather weighed the relative size of boundaries and decided the appropriate grouping based on that information. This could have happened if the subjects were over-briefed during the course of the training session, and understood the exact nature of the task; or if the items were judged as too complex, confusing or tiring to be re-processed each time, and subjects decided to base their decision on what they could tell were cues that varied from one token to the next--the relative boundary size information.

By ignoring all other information, even Overall Bias was prevented from applying, and these items presented a perfect distribution of scores according to the relative size of the two manipulated boundaries, with a cross-over point between interpretations very close to the point of phonetic equality between the two boundaries.

The item *Check In* was tested again with longer intervals, and this time produced a clear and strong First boundary Conditional Bias effect, in line with our predictions about the structure. This suggests that something in the training materials or briefing session for the short *Check In* item was responsible for subjects' choice to ignore the instructions asking them to treat each token as a new and distinct occurrence of the sentence.

The item *Eve or Jude and Sue* instead presented a more complex structure, with varying conjunctions and a disambiguation that relied heavily on the scope of quantifiers *Both* and *Either*, rather than on a scenario or sentence continuation task, which may have exhausted subjects more quickly making them unwilling to process all 49 tokens of the structure.

Lastly, the item *Attack Plan* was reported during our training sessions as having two meanings that were difficult to keep separate given the materials provided, and it is possible that by finding it too difficult to decide which meaning was intended, while at the same time not having a strong bias in favor of either, subjects decided to ignore the processing altogether and focus exclusively on the phonetic cues.

Future work on the subject should verify whether this is a viable explanation, or whether the difference lies in the variability of the phonetic thresholds for Conditional Bias, instead. Furthermore, these findings underscore the importance of developing an independent extension to the experiments, which would test the stimuli's grammaticality, the presence of Overall Bias, as well as the suitability of the task and of the training materials and experimental setup. This would allow experimenters to have more concrete expectations of the results of each item, and to better understand whether the variability across items is due to random noise, or an actual linguistic difference they may have not yet considered.

8.3 The Interaction of Overall and Conditional Bias

Our only original prediction about the interaction of Overall and Conditional Bias was that the latter would affect items with larger boundary sizes than the former, such that the resulting Contour Line in the Distribution Plot would start off as a diagonal, and would then flatten out horizontally or vertically depending on the type of Conditional Bias present.

Only a few items actually displayed this behavior, however, as for several items, the interaction of the two sources of Bias was such that one or the other appeared to be completely obscured. In 5 cases, extremely strong Overall Bias neutralized the Conditional Bias, since all items that the Conditional Bias would have affected already received that interpretation by the Overall Bias. In 9 other cases, the

Conditional Bias was so strong that the Contour Line turned out to be completely horizontal or vertical (such that only one boundary had any effect on the distribution of scores), and we could only assume that the Overall Bias diagonal would have started below or to the right of the Conditional Bias line.

In the latter case, we turned to the MWresponse score of token (0,0) to gain information about the strength of the item's Overall Bias, and to confirm our hypothesis. However, in comparing items tested over both long and short ranges (max 360 or 180 ms lengthening respectively), with a strong Conditional Bias effect, we noticed a Magnification effect on the scores for this item.

For the long version of item *Win Over*, for example, the smaller number of tokens receiving the late break non-Conditional Bias interpretation (which dropped from 13 to 4 because of changes in the experimental setup and the number of tokens run overall as well as the phonetic increments between boundary levels), received a higher score than their counterparts in the short-increments version of the item. It is unlikely, since the sound file as well as the item were identical, that this difference in scores is due to an actual difference in Overall Bias.

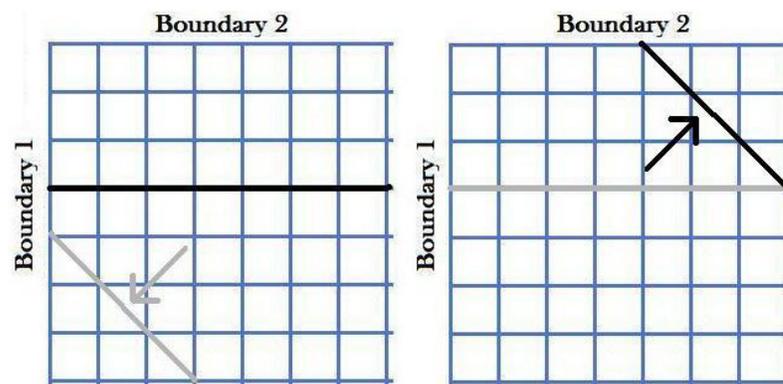


Figure 71: Reciprocally Obscuring Effects of the Interaction of Overall and Conditional Bias

Rather, it seems that this was a byproduct of the experimental setup, such that subjects would try to simplify the form of the demarcation line between the two interpretations, shifting the Overall Bias more towards an early or late break interpretation in order to be obscured by, or totally obscure, the Conditional Bias effect it was interacting with (see Figure 71).

It seems that this effect was caused by the experiment setup, in which subjects were exposed to all tokens within the matrix for each particular item, and as such could keep a mental note of how many tokens received which score, which according to the *Win Over* results seems to affect which method they were using to discriminate between interpretations.

Future research should focus on confirming the presence and determining the extent of the interaction between Overall and Conditional Bias effects, and whether this potential magnification of scores is truly a byproduct of the experimental setup, or is just a further processing shortcut employed by speakers to determine what probability to give each interpretation given the information about relative boundary strength.

8.4 Concluding Remarks

This thesis relates the results of a series of experiments aimed at building and testing a model of ambiguous sentence interpretation based on processing and processing shortcuts. Although we believe that two boundaries contribute information to the disambiguation of the structure, their import is not equal and one tends to gain a processing advantage through a Conditional Bias, whose effects may be magnified by interacting with the sentences' intrinsic Overall Bias.

In addition to supporting and developing our model, the studies in this thesis also provide evidence for the fact that Boundaries much shorter than those

traditionally considered in the literature can have a significant and noticeable effect on meaning determination. We believe these Boundaries do so by providing extra Time at a crucial Point of Disambiguation, such that subjects can process the preceding structure for meaning in a way that will affect the probabilities of the two possible structures that could follow. Depending on the item's structure and the experimental task, this Point of Disambiguation may occur before the ambiguous portion of the sentence has even been pronounced, as in the case of Particle Verbs, Adverbs, or tasks with a forced-pairing format. Furthermore, we present data suggesting that the amount of Time at which Conditional Bias becomes salient may be Universal, as it is constant across most items and structures tested.

These studies also open up a number of avenues for future research, including the domain of application of Conditional Bias, its interaction with Overall Bias, and the effects of the experimental setup on subjects' processing choices within the sentence, which in turn affect our results.

REFERENCES

- Carlson, K., Clifton, C., & Frazier, L. (2001). Prosodic boundaries in adjunct attachment. *Journal of Memory and Language*, 45, 58-81.
- Clifton, C., Carlson, K., & Frazier, L. (2002). Informative prosodic boundaries. *Language and Speech*, 45, 87-114.
- Frazier, L., Taft, L., Roeper, T., Clifton C. Jr., & Ehrlich, K. (1984). Parallel structure: A source of facilitation in sentence comprehension. *Memory & Cognition*, 12, 421-430.
- Gussenhoven, C. (1999). Discreteness and gradience in intonational contrasts. *Language and Speech*, 42, 283-305.
- Gussenhoven, C. (2004). *The phonology of tone and intonation*. Cambridge: Cambridge University Press.
- Kraljic, Tanya and Susan E. Brennan. 2005. "Prosodic disambiguation of syntactic boundaries: for the speaker or for the addressee?" *Cognitive Psychology* 50, p. 194-231.
- Krivokapić, Jelena. 2007. "Prosodic planning: Effects of phrasal length and complexity on pause duration", *Journal of Phonetics*, 35(2): 162-179.
- Ladd, R. (1996). *Intonational phonology*. Cambridge: Cambridge University Press.
- Ladd, R., & Morton, R. (1997). The perception of intonational emphasis: Continuous or categorical? *Journal of Phonetics*, 25, 313-342.
- Lehiste, I. (1973). Phonetic disambiguation of syntactic ambiguity. *Glossa*, 7(2), 107-122.
- Lehiste, I., Olive, J. P., & Streeter, L. A. (1976). The role of duration in disambiguating syntactically ambiguous sentences. *Journal of the Acoustical Society of America*, 60(5), 1199-1202.

- O'Malley, M. Kloker, D. & B. Dara-Abrams. (1973). "Recovering parentheses from spoken algebraic expressions" *Audio and Electroacoustics* 21(3), 217- 220.
- Pijper, Jan R. de, & Sanderman, A. A. (1994). On the perceptual strength of prosodic boundaries and its relation to suprasegmental cues. *Journal of the Acoustic Society of America*, 96, 2037-2047.
- Price, P.J., S. Ostendorf, S. Shattuck-Hufnagel & C. Fong. 1991. "The use of prosody in syntactic disambiguation" *Journal of the Acoustical Society of America* 9, pp 2956-2970.
- Remijnsen, B., & Heuven, Vincent J. van. (2003). On the categorical nature of intonational contrasts. *The phonological spectrum* (pp. 225-246). Amsterdam: John Benjamins.
- Schafer, A., Speer, S. A., Warren, P., & White, S. D. (2002). Intonational disambiguation in sentence production and comprehension. *Journal of Psycholinguistic Research*, 29(2), 169-182.
- Scott, D. A. (1982). Duration as a cue to the perception of a phrase boundary perception. *Journal of the Acoustical Society of America*, 71, 996.
- Snedeker, Jesse & John Trueswell. 2003. "Using prosody to avoid ambiguity: Effects of speaker awareness and referential context", *Journal of Memory and Language*, Vol 48: 103-130
- Streeter, L. A. (1978). Acoustic determinants of phrase boundary perception. *Journal of the Acoustical Society of America*, 64, 1582-1592.
- Watson, D., & Gibson, E. (2004). The relationship between intonational phrasing and syntactic structure in language production. *Language and Cognitive Processes*,
- Wightman, C., Shattuck-Hufnagel, S., Ostendorf, M., & Price, P. (1992). Segmental durations in the vicinity of prosodic phrase boundaries. *Journal of the Acoustical Society of America*, 92, 1707-1717.