A CROSS-LINGUISTIC STUDY ON SYNTACTIC AND DISCOURSE BOUNDARY CUES IN SPONTANEOUS SPEECH

DISSERTATION

Presented in Partial Fulfillment of the Requirements for
the Degree Doctor of Philosophy in the Graduate
School of The Ohio State University

By

Yee-Jean Janice Fon, M. A.

*****

The Ohio State University
2002

Dissertation Committee:

Professor Keith Johnson, Advisor

Professor Mary E. Beckman

Professor Shari Speer

Approved by

Adviser
Department of Linguistics

ABSTRACT

This study focuses on the prosodic and acoustic-phonetic cues at discourse and syntactic boundaries in divergent languages—English, Guoyu, Putonghua, and Japanese. Speech was elicited by having talkers describe the events in *The Pear Story* film. Recorded data were transcribed and segmented into discourse and syntactic units. Prosody was labeled following the conventions of Tones and Break Indices in each language. Acoustic-phonetic measurements of F_0, syllable duration, and syllable onset intervals (SOIs) were taken on the digitized data. A comparison of different dimensions of data—discourse/syntax, acoustics, and prosody, was made in order to examine boundary cues in speech.

Results showed that both language-universal and -specific cues exist. Prosodically, structural boundaries are indicated by intonation phrase (IP) boundary breaks in English and Japanese while in Guoyu and Putonghua, they can be indicated by a break level that is of a minor phrase boundary or higher. For all four languages, the proportions of IP breaks reflect structural hierarchy in a positive manner, although the degree of reflection changes with language.

Acoustic-phonetic cues show more cross-linguistic variations. Pitch reset is a prevalent cue for structural boundaries in all languages but English. For the three languages that show consistent pitch reset and declination across structural boundaries, the magnitude is reflective of structural hierarchy. Final lengthening of boundary syllables and SOIs is the most universal cue for signaling structural boundaries, and the degree of final SOI lengthening is reflective of disjuncture hierarchy. All four languages examined in this study showed similar patterning. However, there are also language-specific cues. In English, in addition to final lengthening, initial pitch-accented syllables and SOIs are also lengthened. In Guoyu and Putonghua, the scope of final syllable lengthening is widened to include the penultimate syllable. English and Putonghua are similar to each other in that there is no reflection of hierarchy in the degree of final syllable lengthening. On the other hand, Guoyu and Japanese are more alike in this regard since both languages show some reflection of hierarchy through the degree of boundary syllable lengthening. Bigger structural boundaries are signaled by a smaller degree of lengthening.

Dedicated to my parents

ACKNOWLEDGMENTS

Like many dissertations before me and many yet to come, this study could not have been possible if it were not for the help of many, many angels in my life. Unlike in the biblical era, where angels were robed in white and flew with wings, angels in the modern age come in many disguises—as advisors, mentors, coworkers, friends, family members, or even people I do not personally know. In the following, I wish to do my best to show my appreciation to each one of them.

First of all, I would like to thank my mentor and my advisor, Keith Johnson, for intellectual support and guidance. Keith has been my advisor ever since I came to The Ohio State University. Throughout the five years, he has always been patient with my struggles and giving me helpful and creative advice in a way that I can handle. I especially want to thank him for his tolerance to many of my crazy ideas and bizarre thoughts regarding the direction of my study and experimental designs. If it were not for his support and nonjudgmental attitude, I probably would not have stayed in the field of linguistics anymore, let alone finishing this dissertation.

I would also like to thank my other mentor, Mary Beckman. I have always enjoyed Mary's classes and lectures. She has a way of describing phonetic principles that makes them much more vivid and interesting than it is in books. Despite her busy

schedule, she has always been generous with her time and comments. Her insightful thoughts and suggestions have been a stimulating challenge to me. Although it is sometimes embarrassing to be caught red-handed with a blatant error, I am still glad such moments exist, because it means I am another step closer to perfection.

I thank Shari Speer, my other committee member, for suggestions on experimental designs and comments on results and interpretations. She provides me with a psychologist's perspective, one that I lack from my training, which helps me develop a well-rounded view of language.

Since this is a cross-linguistic project, and since I am only a native speaker of one of the languages in this study, I am grateful to many of the people that helped me out throughout the process. I thank Manuel Diaz for kindly lending me the videotape of *The Pear Story* and providing me with tips to recording. I also thank Matthew Hyclak for setting up the recording environment. I would like to thank my native speaker experimenters for recording subjects: Keith (again!) and Terah Schamberg for English, Tsan Huang for Putonghua, and Kiyoko Yoneyama for Japanese. I am also thankful to my transcribers: Terah (again!) for English and Shirai Aya and Emma Vrcek for Japanese. My ToBI labelers also deserve my deep thanks, as anyone who does prosodic labeling knows how painstaking and demanding the process can be: Craig Hilts for English, Tsan Huang for Putonghua, and Jennifer J. Venditti for Japanese. As a phonetician and psycholinguist by training, I would also like to thank the syntacticians that helped me out with the syntactic analyses: Martin Jansche for Mandarin, and Akihiro Kano and Mineharu Nakayama for Japanese. I would also like to thank Chao-Feng Wu for helping me key in the data and Kuenhi Tsai for statistical analyses.

I am especially thankful to all my subjects. Many of them are friends or friends' friends. Without them providing useful data, I could not have had such interesting results.

I also wish to thank the National Institute of Development and Communication Disorders for funding the project (NIDCD grant no. R01DC04421), and thank Keith for hiring me as a research assistant. If it were not for the funding that alleviates me from my teaching load, I could not have finished the project so quickly and smoothly.

I am also thankful for the moral support of my labmates, especially Kiyoko Yoneyama, Matthew Makashay, Tsan Huang, and Peggy Wai-Yi Wong. They are always there to listen to my complaints and share my worries. I would also like to thank the Psychoid meeting group for brainstorming, experience sharing, and problem solving. I learned a lot about the ethics of running human subjects from many of the members, especially Laurie Maynell and Allison Blodgett.

As I am only a naïve computer user, and computers are well known for mischievously taking advantage of low-tech people like me, I would like to express my deep thanks to my good friend Pang-Hung Hsu for the technical support. He spent numerous late nights and weekend hours troubleshooting and solving computer problems for me. He also provided me with the warm mental support that I needed throughout the ups and downs in the past five years. In a way, he is more like a family than a friend to me, and I thank him for that.

I would also like to express my thanks to my Taiwanese gang. Being thousands of miles away from home has not been easy for a home person like me. I am especially grateful to have a group of friends from home, friends to laugh together and gossip with, and friends to provide care, food, and even words that are of home atmosphere. They are Ya-wen Chang, Hui-Chen Chen, Chi-Min (Jason) Cheng, Pi-Yu Wang, Hui-Ying Hung, Pang-Hung Hsu, Jer Lai Kuo, Kang Earnest Liu, Wai Man Raymond Liu, Si-Yi Oh, Jackie Yu-Lin Shao, Mei-I Su, Ping-Hui Tseng, and Chao-Feng Wu. Without them, I might still have somehow survived, but would never be as happy and content.

Finally, I would like to give my deepest thanks to my family, including my parents, my godparents, and my brothers. I thank my father and mother for raising me and providing me with a home full of love and care. I thank my godparents, especially my godmother, for taking me in and adopting me after my mother died. I also thank my stepmother for taking care of my father while I am far away from home so that I am less worried about his well-being. I would also like to thank my brothers, Cary, Michael, and Andy, who accept me as who I am, no matter how much or how little I have accomplished.

I distinctly remember an article entitled *Thank God*, written by Chi-Fan Chen, a Hong Kong physicist and a writer, that I studied as a high school student. It talked about how people turn to thank God because there are just too many people to give thanks to. Although writing a dissertation and earning a Ph.D. degree is nowhere near what forerunners in this field have accomplished, it is a personal feat for me, coming from a small village where going to college used to be a rare luxury. Naturally, this feat does not only belong to me. It also belongs to my friends and my family, and all those that helped me directly or indirectly throughout this process. Like what Chen has mentioned in his article, there are too many angels in my life that I need to give my thanks to, and I could not possibly give to them all. Therefore, I would like to give my deepest and most sincere thanks to God, who kindly sent these angels to me.

VITA

November 13, 1971	Born – Taipei, Taiwan
1993 – 1997	Research Assistant, National Taiwan University
1994	B. A. Foreign Languages and Literature, National Taiwan University
1997	M. A. Linguistics, National Taiwan University
1998 – present	Graduate Teaching and Research Associate, The Ohio State University

PUBLICATIONS

Research Publication

1. Cheung, H. & Fon, J. (2002). The construction of classifier system in Mandarin Chinese. Proceeding of 1st Cognitive Linguistics Conference, 425–439.
2. Fon, J., & Johnson, K. (2000). Speech timing patterning as an indicator of discourse and syntactic boundaries. *Proceeding of 6th International Conference on Spoken Language Processing*, *2*, 555–558.

3. Fon, J. (1999). Speech rate as a reflection of variance and invariance in conceptual planning in storytelling. *Proceeding of 14th International Congress of Phonetics Sciences*, 663–666.
4. Fon, J., & Chiang, W.-Y. (1999). What does Chao have to say about tones? —-a case study of Taiwan Mandarin. *Journal of Chinese Linguistics*, *27*, 1, 15–37.
5. Fon, J. (1998). Variance and Invariance in Speech Rate as a Reflection of Conceptual Planning. *Proceeding of 5th International Conference on Spoken Language Processing*, *7*, 3095–3098.

FIELDS OF STUDY

Major Field: Linguistics

TABLE OF CONTENTS

Page

Appendices:

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Background

The vastness of input one receives everyday tends to go unappreciated. Every moment, the nerve system is bombarded with myriads of information, from lower level signals such as smell, taste, sight, sound, and touch, to higher level signals such as facts, concepts, arguments, beliefs, and emotions. How exactly human beings process incoming information simultaneously and successfully is still a question not satisfactorily answered. However, it is believed that segmentation and grouping abilities play a major role in facilitating processing and cognition.

Take visual input for example. In order to explain how human brains use 2-D information that falls onto the retina to reconstruct 3-D images, Marr (1982) proposed a three-staged computational model for visual cognition—the primal sketch, the 2½-D representation, and the 3-D representation. The primal sketch is the first pass of visual processing. It takes in raw physical signals such as light intensity in the visual world, computes light intensity changes, and assigns edges to visual objects where

zero-crossings occur. In other words, the primal sketch is performing a rough segmentation of visual objects using very simple physical signals.

In the audio world, segmentation is also very important. Unlike written English, and many other written languages using phonetic alphabets of some kind, where words are conveniently separated by spaces, sentences by periods, and paragraphs by paragraph breaks, spoken languages are more like Chinese writings, where no apparent word boundary cues are provided, and sometimes even more like ancient Chinese writings, where both word and sentential boundaries are left unmarked.

Nonetheless, human beings seem to process auditory information with graceful ease. An average listener hardly has any problem identifying words, sentences, and even topic boundaries upon hearing a stretch of continuous speech. Although one might argue that through word identification and lexical access, the segmentation problem is virtually nonexistent, much is still left unexplained. First of all, there is no one-to-one mapping between pronunciation of a word and the word itself. Table 1.1 shows the 28 most frequent pronunciations of the word *and* and their token frequencies in the Buckeye Speech Corpus, which is a corpus of spontaneous speech consisting of 300,000 words from recorded interviews of 40 speakers (Raymond et al., 2001). Notice that the first three pronunciations listed in the *Merriam-Webster Online Collegiate Dictionary* (2002), [ənd], [ʌnd], and [ænd] ([d]can be omitted in all three) are not even the most frequent. If more than two dozens of pronunciations can be associated with a simple *and*, the order of complexity would be astronomical for any sentence of average length. Even if all of the tokens are stored in one's mental lexicon under the entry *and*, this segmentation-by-lookup method would still not be as efficient as it needs to be.

Token	Frequency	Token	Frequency	Token	Frequency	Token	Frequency
ɛn	0.237	ænd	0.023	æ	0.003	ənd	0.002
n̩	0.166	ɛnd	0.017	ə	0.003	əŋ	0.002
ən	0.155	æɾ̃	0.013	ɛm	0.003	ɛænd	0.002
æn	0.144	ɛɾ̃	0.013	ɛŋ	0.003	n̩n	0.002
ɪn	0.083	əɾ̃	0.009	d	0.002	ɪnd	0.002
ɨn	0.030	ʌɾ̃	0.008	ən̩	0.002	ɪɾ̃	0.002
n	0.029	m̩	0.005	əm	0.002	ɪ	0.002

Table 1.1: The 28 most frequent pronunciations of *and* and their token frequencies in the Buckeye Speech Corpus adapted from Raymond et al. (2001).

Secondly, people with little or no mental lexicon also seem to be capable of segmenting speech, albeit in perhaps a rudimentary fashion. Studies have shown that infants as young as 7½ months old can segment out at least some words in a sentential context (Jusczyk, Houston, & Newsome, 1999). In addition, anecdotal stories on non-speakers of a language being able to tell where a sentence ends and another begins are not uncommon.

What is suggested here is not that mental lexicon does not play a role at all in word segmentation (and as a consequence, in sentential and discourse segmentation), but that it may not be the sole actor on stage. Boundary cues comparable to the primal sketch in Marr's (1982) computational visual cognition model should also be present somehow, somewhere in the auditory domain in order to account for the speed and accuracy in speech segmentation demonstrated by just any human listener. In other words, the analogy of comparing spoken language with Chinese writing might be proven

inappropriate if one looks harder into the signal. Punctuation marks in the audio domain might very well exist, although the cues might more likely be degraded as compared to their written counterparts.

1.2 Significance

This study intends to look at boundary cues at discourse and sentential levels using a small multi-lingual spontaneous speech corpus. Although read speech is more controlled and easier to handle, the richness and dynamics of spontaneous speech cannot be paralleled. Considering the difference in the likelihoods of encountering spontaneous versus read speech in one's everyday life, one is more likely to find segmentation cues that are actually present in online speech processing using spontaneous speech. Moreover, segmentation strategies might also differ between the two genres since cues in read speech are likely to be more dependent on written forms and punctuations rather than spoken grammar.

Another significance of this study lies in its cross-linguistic perspective. Segmentation literature, as in any other literature in the language studies, is dominated by English data. Although by modus ponens, what is true in English should also be true of Language, what is not true of English may or may not be true of Language. The only real fault-proof way to verify what Language is is to examine every single language in the world, which is of course too time- and energy-consuming to be empirically possible. A second best option would be to look at a range of languages, especially those that are not as genetically related. This is especially important if one is interested in language-universality. Studies based on only one language or a couple of closely related languages are not likely to tease apart universals from specifics. By looking at languages from different families—English, Japanese, and Mandarin, this study intends to investigate both ends of the spectrum. The universal end is especially important if the cues found are intended to explain how infants and non-speakers perform rudimentary segmentation. On the other hand, the specific end is important in that it can account for the discrepancy in speed and accuracy between native and non-native speaker segmentation, ceteris paribus.

The three languages were chosen because they represent a continuum of tonal target specification densities, with English being the sparsest and Mandarin being the densest. Two varieties of Mandarin were included in the study—Guoyu (i.e., Taiwan Mandarin) and Putonghua (i.e., Beijing Mandarin). These two varieties are interesting because among many things, they differ in the existence of lexical stress. Guoyu has almost no lexical stress while Putonghua does. For example, in Putonghua, *dong1-xi1* denotes 'east-west' while *dong1-xi0* denotes 'thing, stuff'.[1] On the other hand, the two lexical items are homophones (both as *dong1-xi1*) in Guoyu due to loss of stress contrast. By including both languages of different families and dialects of the same language, this study intends to look at language universality and specificity and how they interact with language relatedness.

[1] The romanization of Pinyin system is adopted here. The numbers after each syllable indicate tonal categories. There are four tones in both varieties of Mandarin (Chao, 1968). Tone 1 is high-level, Tone 2 is mid-rising in Putonghua and mid-dipping in Guoyu (Fon & Chiang, 1999), Tone 3 is low-dipping, with an allotone of low-falling, and Tone 4 is high-falling. Tone 0 indicates a neutral tone, which is designated for unstressed syllables.

Finally, instead of focusing on word segmentation, as the majority of literature did, segmentation at sentential and discourse levels is the main interest of the current study. Although word segmentation is important and interesting in its own right, sentential and discourse segmentation is also important if not more, because what one encounters most often in his everyday linguistic life is sentences and discourses, not words in isolation. How the organization of discourse and syntax is indicated in spoken language cannot possibly be fully answered by mere addition or amplification of word segmentation. This study thus intends to contribute somewhat and somehow to the relatively scanty literature of segmentation in a larger domain, and by doing so, hopes to gain more insight into how an organism as complex as *Homo sapien* think, learn, and speak.

1.3 Specific Aims

Although there can be literally myriads of hypotheses concerning what cues to look for in syntactic and discourse segmentation, this study focuses itself only on prosodic and acoustic-phonetic cues. This is inspired by the parsimonious principle in Marr's (1982) visual cognition model. If cues as basic as changes in light intensity can be used as a basis for segmenting objects in the visual domain, there is no reason why simple prosodic and acoustic-phonetic cues could not be utilized for segmenting speech streams in the audio domain. Moreover, language universal cues should by definition be simple to detect since they are meant to ease the acquisition process of infants. Signals that involve complex linguistic knowledge have to be learned through extended exposure and do not fulfill such criteria. Acoustic-phonetic cues are chosen as candidates because they can be detected and perceived as long as listeners have normal hearing, at least in most cases. Prosodic cues are also chosen because studies have shown that they are one of the very first linguistic elements acquired. Although the exact starting date of acquisition is hard to pinpoint, Nazzi, Bertoncini, and Mehler (1998) showed that French newborns can discriminate languages that are from different rhythmic classes (e.g., British English and Japanese) but not those that are from the same class (e.g., British English and Dutch). In other words, language acquisition, or at least prosody acquisition, may start as early as during the gestation period.

This study aims to investigate three effects in the prosodic and acoustic-phonetic cueing of sentential and discourse boundaries. First of all, one would like to examine whether there is a boundary effect. That is, whether discourse boundaries are encoded prosodically and acoustically in spontaneous speech. Realization of a boundary effect would be that signals at sentential and discourse boundary positions are different (usually magnified) compared to those within the boundaries. The second effect to be examined is the hierarchy effect. That is, whether discourse boundary strength is mirrored by prosodic and acoustic strength. Realization of such an effect would be a (usually positive) correlation between boundary strength and prosodic and acoustic strength at the boundary position. Finally, the third effect to be examined is the language effect. In other words, whether encoding is different in prosodically different languages. Realization of this effect would be that boundary signals are patterned differently according to different languages.

1.4 Organization

This dissertation is organized into eight chapters. Chapter One is the introduction. It provides the background of the issue of interest, the significance, and the specific aims of this study. Chapter Two consists of the literature review. This includes segmentation literatures with regards to realizations of boundary and hierarchy effects in different languages, especially at the sentential and discourse levels. Chapter Three describes the methods used in the experiments of this study. This includes recording paradigms and data analyses. Chapter Four reports the results of prosodic breaks as a cue in discourse and sentential segmentation. Chapters Five and Six report the results of the two acoustic-phonetic parameters examined in the study, pitch and duration. Chapter Seven discusses the results in Chapters Four through Six and the significance of such results. Finally, Chapter Eight summarizes the study and attempts to relate the results of the study to the segmentation literature.

CHAPTER 2

LITERATURE REVIEW

Three hierarchical organizations are at work in a spoken discourse setting. There is the structural hierarchy that follows the morpho-syntactic tradition of segmenting discourse into sentences, the highest level of hierarchy within the organization. Sentences are then segmented into clauses, clauses into phrases, phrases into words, and words into morphemes. There is also the pragmatic-discourse hierarchy, which segments discourse into hierarchically organized discourse units of various sizes. The organization of discourse is also viewed as structural, although the definition of discourse units and thus the architecture vary widely with frameworks. The third hierarchical organization is concerned with prosody. Among the theories of prosodic structure, the one by Selkirk (1984, 1986) is the most influential. In this framework, the highest level is an utterance. An utterance is then segmented into intonational phrases, intonational phrases into phonological phrases, and phonological phrases into prosodic words.

As this study is mainly interested in sentential and discourse boundary cues in prosodic and acoustic-phonetic domains, and how hierarchy cues are manifested in discourse organization, what is of concern here is whether there is a consistent mapping

between the structural (both morpho-syntactic and pragmatic-discourse) and prosodic hierarchy. By consistent mapping, one does not imply that there is a rigid one-to-one correspondence. Rather, the predictions are that prosodic and structural boundaries tend to overlap to provide boundary cues, and prosodic hierarchy correlates with structural hierarchy in a positive manner to provide hierarchy cues. Since prosodic hierarchy, like structural hierarchy, is constructed with universality in mind, most of the cross-linguistic differences should be absorbed by acoustic-phonetic cues, and only few would be realized in the prosodic domain.

This chapter is divided into two sections. The first section is devoted to how clausal and sentential boundaries are signaled in speech. Clauses are often considered as the basic syntactic unit in spontaneous speech studies. Unlike written texts, where sentences are explicitly marked by some kind of punctuations (e.g., '.' in English and '。' in Mandarin and Japanese), the definition of sentence in a spoken discourse is far from clear. Therefore, clause is oftentimes used as a substitute unit for sentence in discourse analysis since the definition of a clause is more self-contained than a sentence. A clause is defined by the elements it contains (i.e., a verb or a verbal construct), rather than its relationship with neighboring clauses or other structural units. Although this study is mainly concerned with discourse hierarchy, clausal boundaries are also relevant because they are potential locations for discourse boundaries. Discourse boundaries are always clausal boundaries, although the other way around is not necessarily true. By comparing clausal boundaries that either does or does not coincide with discourse boundaries, one can have a clearer understanding of how boundary cues are patterned in a spoken discourse.

The second section of the chapter focuses on hierarchy cues, especially those of discourse boundary. There are two levels of structural hierarchy that are of interest in this study. One is the hierarchy between clausal and discourse boundaries. The latter is deemed as higher in hierarchy than the former. The other is the hierarchy within the discourse framework. That is, differences in boundary size between a big versus a small discourse unit might somehow be reflected by prosodic and acoustic-phonetic cues. In this study, the boundary of a big discourse unit will be considered as the high end of a continuum of discourse boundary sizes, while that of a potential discourse boundary (i.e., a clausal boundary) will be the low end. A small discourse unit will be deemed as representative of a boundary size that is somewhere in the middle.

2.1 Boundary Cues

Although there is not a one-to-one correspondence between prosodic and structural boundaries, many studies have shown that the two coincide with each other. Price, Ostendorf, Shattuck-Hufnagel, and Fong (1991) examined the relationship between the two organizations. In order to quantify prosodic boundaries, they devised a 7-scaled prosody labeling system to label boundary strengths: 0—clitic group; 1—prosodic word; 2—accentual phrase; 3—intermediate phrase; 4—intonational phrase, 5—breath group; and 6—sentence.[1] Professional announcers recorded a range of sentences that are ambiguous (e.g., *Mary knows many languages, you know* vs. *Mary knows many languages you know*). Subjects were played the sentences and were asked to determine which version they belonged to and give confidence ratings. Prosodic analyses of the recordings showed that presence of an intonational phrase boundary (break index 4 or higher) is often a reliable cue for an embedded or conjoined clausal boundary. Relatively larger break indices tend to indicate high syntactic attachments. Prominence plays mainly a supporting role, and is the sole cue in only a few sentences. Perception results showed that subjects were very accurate in sentence discrimination (84% correct), although their confidence judgment was low (52% confident). Prosodic boundary cues can often be used to disambiguate the two interpretations. The authors thus concluded that clausal boundaries containing complete sentences nearly always coincide with boundaries of major prosodic constituents. Naïve listeners can thus reliably use prosody to separate structurally ambiguous sentences, and use prosody to assign syntactic structures.

When prosodic and structural boundaries are in conflict, comprehension seems to be impeded. Speer, Kjelgaard, and Dobroth (1996) examined how temporary syntactic closure ambiguities can be resolved by prosodic structures using an end-of-sentence comprehension task and a cross-modal naming task. In the first task, subjects were asked to press the button when they understood the sentence. Stimulus conditions include sentences with either cooperating or conflicting prosody: (1) cooperating prosody, late closure: e.g., [*Whenever the guard checks the door*] % [*it's locked*]; (2) cooperating prosody, early closure: e.g., [*Whenever the guard checks*] % [*the door is locked*]; (3) conflicting prosody, late closure: e.g., [*Whenever the guard checks* % *the door*] [*it's locked*]; (4) conflicting prosody, early closure: e.g., [*Whenever the guard checks*] [*the*

[1] The lower part of the hierarchy (0–4) is similar to what Selkirk (1984, 1986) proposed in that it represents the abstract structure marked by licensing of boundary tones and other segmental/tonal features, and also by phonetic cues such as final lengthening and accent-related strengthening. However, break indices 5 and 6 are rather different in that they represent a subjective sense of a discourse hierarchy from continuous variation in pitch range relations, degree of final lowering and lengthening, etc. In other words, the former represents prosodic boundaries and categorical markers to these boundaries while the latter represents the gradient acoustic variation of things at different prosodic positions/levels.

door % is locked]; (5) baseline prosody, late closure: e.g., [*Whenever the guard* checks the door*] [*it's locked*]; (6) baseline prosody, early closure: e.g., [*Whenever the guard* checks*] [*the door is locked*].[2] In the second task, sentences were chopped at the word *door* and the next word was shown as a visual input. Subjects were asked to say the word as quickly as possible. Similar results were found in both tasks. In the cooperating condition, there is no difference in reaction time (RT). In the conflicting and the baseline conditions, late closure has shorter RT than early closure, revealing the syntactic bias. However, RT is slower for the conflicting condition. The RT pattern still holds when intermediate phrase boundaries were used instead of intonational. Speer et al. thus claimed that the influence of prosodic structure is not limited to sentences that contain highly salient acoustic boundaries (as in intonation phrase boundaries). When the boundaries of syntactic and prosodic constituents coincide, prosodic boundaries facilitate syntactic parsing, especially for the less-preferred interpretations (i.e., the early closure sentences). When the two are in conflict, prosody will show an interfering effect to the syntactic parser.

There is also a long tradition in looking for clausal and sentential boundary cues in the acoustic-phonetic domain. People have been searching for them ever since the 1970s. Over the years, many cues have been proposed, tested, and confirmed. Among them, final declination and pitch reset are proposed as good boundary indicators. Cooper and Sorensen (1977) examined the relationship between fundamental frequency contours and clausal boundaries using read sentences. Clausal boundaries between conjoined main clauses, between a main and an embedded clause, and between a preposing element and

[2] '%' indicates an intonation phrase boundary, brackets indicate syntactic groupings, and '*' indicates a high pitch accent.

its main clause were examined. Final declination and pitch reset were consistently observed within and across a clausal boundary, respectively, and there is no gender difference in the pitch contour patterning aside from absolute pitch height.

Fisher and Tokura (1996) showed that declination is also prevalent in English and Japanese motherese (*N* = 3, 14-month-olds for both languages). In both languages, pitch excursion is more drastic at utterance-final syllables. This is also true of children's speech. Observing children between ages 12 and 20 months for nine months, Snow (1994) found that for accented syllables, English-learning children showed greater tonal excursions at utterance-final than non-utterance-final positions.

Another reliable cue indicating clausal disjunctures that has been studied extensively is the silent pause. Using both read and spontaneous speech (i.e., a radio talk), Goldman-Eisler (1972) showed that silent pause serves a demarcation function at structural disjunctures in English. In both genres, a correlation between the existence of a silent pause and clausal and sentential boundaries was found.[3] However, not all clausal disjunctures are indicated by silent pause. Of the three types of clausal boundaries examined—between matrix and relative, between matrix and subordinate other than relative (e.g., *if*-, *because*-, *since*-, *while*-clause, etc.), and between two matrix clauses, disjuncture between coordinate clauses is the most likely to be indicated by silent pause while that between a matrix and a subordinate is the least likely.

Fisher and Tokura (1996) also confirmed this finding by showing that silent pause is a good boundary indicator in English and Japanese motherese. 96% of pauses in English and Japanese coincided with utterance-final boundaries and 59% of

[3] Sentential boundaries are possible in this study because radio talk, even if unscripted, is usually thoroughly planned, and therefore more similar to text.

grammatically distinct, utterance-final boundaries in English and 69% of grammatical boundaries in Japanese were followed by pauses.

Infants actually also prefer silent pause at a clausal boundary to that within. Kemler Nelson, Hirsh-Pasek, Jusczyk, and Cassidy (1989) played speech utterances of child-directed speech (CDS) and adult-directed speech (ADS) with 1-sec pauses inserted within (i.e., non-coincident) or between clauses (i.e., coincident) to 8½-month-olds. A head-turning paradigm was used to test infants' preferences. Results showed that longer duration of head orientation was induced when infants heard CDS with coincident than with non-coincident segmentation cues. This difference did not show up with ADS. Coincident CDS also induced longer duration of head orientation than coincident ADS. But no such difference showed up between the non-coincident conditions. Kemler Nelson et al. thus suggested that CDS provides cue packaging of syntax to which infants attend while ADS does not. In addition, infants are more likely to attend to utterances where cue packaging is consistent. This preference is advantageous in acquisition since acoustic-phonetic boundary cues are strongly linked to syntactic boundaries in CDS.

Another cue that has been proposed as indicating sentential boundary is the final lengthening effect. This effect describes the observation that units at sentence-final position tend to be lengthened. Lehiste (1975b) examined one such unit—foot duration. Tetrametrical sentences such as *Never visit busy cities* were recorded. Lehiste measured foot duration and found that at sentence-final positions, foot duration is lengthened and duration characteristic of the foot type at non-sentence-final positions is neutralized.

Oller (1973) studied the lengthening effect using a smaller unit—the segment. Nonword syllables such as *bab* was recorded in a carrier sentence such as *Say a* _____.

Results showed that there is a lengthening effect at the sentence-final position on both consonants (i.e., *bab*) and vowels (i.e., *bab*) regardless of stress (i.e., primary-stressed vs. unstressed) and intonation patterns (i.e., imperative, declarative, and interrogative). Using a phonetically balanced read corpus, Campbell and Isard (1991) also showed that sentence-final syllables have more lengthened peaks and codas as opposed to onsets using normalized duration as the dependent measure. Thus, in English, lengthening in pre-pausal syllables is more confined to the rhyme than the whole syllable.

Shen (1992) used famous utterances of standing ambiguity in literary Beijing Mandarin to demonstrate that both silent pause and final syllable lengthening are robust cues in signaling sentential boundary location. One set of examples is shown below and the two readings are given in (1) and (2). The brackets indicate where the sentential boundaries are and the underlined syllables are the ones she measured for duration. Results showed that syllable duration is significantly lengthened when it is at the sentential boundary position. For example, the *ke* 'guest' in (2) occupies about 13–14% of total speaking time (i.e., excluding silent pause duration) of the sentence while the same word in (1) only occupies about 9–11%. Silent pause always follows a sentential boundary.

(1)

[*xiayu*	*tian*]	[*liu*	*ke*	*tian*]	[*liu*	*wo*	*bu*]	[*liu*]
[rain	day]	[stay	guest	day]	[stay	I	no]	[stay]

It's a rainy day. It's a day (to ask) guests to stay. (Do you want to invite) me to stay or not? Stay.

(2)

[*xiayu*]	[*tian*	*liu*	*ke*]	[*tian*	*liu*]	[*wo*	*bu*	*liu*]
[rain]	[heaven	stay	guest]	[heaven	stay]	[I	no	stay]

It's raining. Heaven (is asking) guests to stay. Heaven (wants you) to stay. I do not (want you) to stay.

Fon and Johnson (2000) proposed that syllable onset intervals (SOIs), the interval between the onset of one syllable and that of the next, which is essentially a combined measure of final lengthening and pause, is a good indicator of clausal boundary. Using a small Guoyu corpus of spontaneous monologs, they found that clause-final SOIs are always lengthened while SOIs in clause-initial and -medial positions do not show any difference in duration. Fon and Johnson thus suggested that Guoyu has essentially isochronous syllable timing. This regular rhythmicity is interrupted primarily at structural boundaries by final SOI lengthening.

By comparing child-directed speech (CDS) and adult-directed speech (ADS) in English, Ratner (1986) showed that the final lengthening effect at clausal boundaries is instrumental to segmentation in language acquisition. Conversations between mothers and their children at different stages (i.e., prelinguistic, one-word, and multi-word), and between mothers and the experimenter were recorded, and vowel durations of matching words in both situations were measured. Results showed that clause-final vowels were longer than clause-medial vowels and the degree of lengthening is a function of listener's age. Lengthening is the most prominent in CDS for prelinguistic stage infants as compared to that for one-word stage and multi-word stage children and ADS for adults.

Similar evidence was also found by Fisher and Tokura (1996). They showed that utterance-final vowel lengthening is a robust cue in both English and Japanese CDS to 14-month-olds.

Snow (1994) provided a piece of evidence that children acquiring English do actively learn final lengthening at utterance-final position. Nine children with a productive vocabulary between 30 and 70 words were observed over a 9 months' period. Accented syllables containing segments of the same type (e.g., *cup* vs. *kitty*) were used for comparison. Results showed that final syllable lengthening appears as a U-shaped trend. It first appears, and then children become isochronous around the transition phase between one-word and two-word stages. Finally, it reappears. Children develop a consistent utterance-final versus -nonfinal contrast in duration within 3 months after they have begun to combine two words into phrases. In other words, an adult-like precursor of final lengthening is present at the one-word period. However, final lengthening at this stage can be accounted for by imitation or passive physiological relaxation. On the other hand, final lengthening after syntax (i.e., two-word stage and after) may need to be learned. Children's tendency to minimize syllable-timing distinctions during the transition stage may represent a period of reorganization when they begin actively controlling features of speech timing that had been passively controlled at an earlier stage.

Although amplitude is a cue that is not often studied, Fisher and Tokura (1996) claimed that amplitude can be a consistent cue for utterance boundaries, although languages may differ in terms of patterning. Looking at English and Japanese motherese, Fisher and Tokura found that at utterance-final positions, English shows a slight increase in amplitude while Japanese a slight decrease.

Most of the boundary cues shown above are very robust. However, they would be less meaningful if listeners cannot perceive these cues effectively. Berkovits (1984) tested English and Hebrew bilinguals and showed that listeners can perceive sentence-final cues regardless of the language source. Sentences from both languages were used and the first part of the sentences was spliced out (e.g., *They gave up the search.* vs. *They gave up the search after three hours*). English- and Hebrew-dominant subjects were asked to determine whether the sentence was finished or not. Results showed that subjects of both groups were able to respond correctly most of the time. The only difference was that dominant languages were detected faster than non-dominant languages and Hebrew finished sentences were more likely to be categorized as unfinished by English-dominant speakers. The author attributed this to the prominent final lengthening effect in English that Hebrew lacks. A second experiment used the same set of sentences (i.e., sentences with finished and unfinished intonation), but with preceding contexts given. Subjects were asked to press the button as soon as they finished reading. Results showed that sentences with sentence-final boundary cues elicited faster RT than sentences without. Whether subjects noticed the intonation differences did not matter. The results suggested that listeners perceive acoustic and prosodic cues at sentential boundaries. Despite slower processing time in the weaker language, perception of sentence-final cues is not a function of language dominance. Instead, it reflects acoustic properties of the dominant language.

2.2 Hierarchy Cues

Although discourse hierarchy and topic boundaries are most easily identified by semantic and syntactic information (Schaffer, 1984), thus theoretically speaking, listeners can comprehend discourse organization well without prosodic and acoustic information, listeners actually prefer to listen to utterances where prosodic and structural hierarchies coincide (Sanderman & Collier, 1996). More importantly, Swerts (1997) found that prosodic and acoustic-phonetic cues actually do help listeners in determining discourse organization. Using a clustering analysis (Rotondo, 1984), Swerts examined whether Dutch listeners agree upon how to define paragraph boundaries. Subjects were given 12 spontaneous monologs to determine where paragraph boundaries were. The monologs were given either only in text or in text combined with speech. The percentage of subject agreement was taken as a measure of hierarchy. Pauses, pitch resets, and boundary tones were measured and results showed that a linear model fit of *boundary = pause + reset + tone* accounted for 58% of the variance. When the scores for subjects labeling with texts only (i.e., semantic and syntactic information) were also included in the model, the explained variance increased to 85%. Swerts concluded that although textual cues are clearly predominant for the labelers, the contribution of the prosodic and acoustic variables is not negligible. Thus, in the following, studies concerning non-semantic and non-syntactic cues regarding discourse hierarchy are reviewed, which include both prosodic and acoustic-phonetic cues.

Prosodically, pitch-accents are often proposed as a cue for signaling discourse structure. Terken (1984) tried to show this in Dutch, using a task of assembling the front view of a house from ready-made parts to elicit spontaneous monologues. Each instruction is defined as a separate topic. Results showed that when introducing a new topic, speakers nearly always use an accented full expression. Once the topic of an instruction has been introduced, there is a considerable decrease in the number of accented expressions, and the number stays low with subsequent mentions. Recency also affects accentuation. De-accentuation could occur if a referent is mentioned in the previous topic. Terken thus suggested that speakers' judgments regarding availability of a referent is sensitive to the thematic structure of the discourse, which is reflected by accentuation. Appropriate distribution of pitch-accents would help listeners to effectively process resources at each point in the utterance. Using the same corpus, Swerts and Geluykens (1994) also showed that there is a correlation between topical structures and the use of boundary tones. Low-ending tones are associated with instruction finality while high-ending tones are associated with nonfinality.

Traum and Heeman (1997) also confirmed the relationship between boundary tones and discourse organization in spontaneous dialogs in English (Heeman & Allen, 1995). Results showed that in clear transitions (i.e., when no overlap occurred), 95% of the utterances were followed by a boundary tone when it was backchanneling or directly related to the previous speaker's utterance. On the other hand, when utterances were related to the utterances prior to the most recent ones, or when they were not related at all, only about 64–72% of the utterances were followed by a boundary tone. Traum and Heeman thus suggested that boundary tone is a primary cue for discourse hierarchy while acoustic-phonetic cues such as pause is secondary.

Since hierarchy implies gradation, it is often realized in the continuous acoustic-phonetic domain. One type of cues that has been proposed is declination and pitch reset. As a pioneer in this field, Lehiste (1975a) showed that listeners can make use of pitch information to determine the positioning of a sentence in a paragraph. She recorded three sentences in isolation and in all six possible orders. The sentences were then spliced out in isolation and played to subjects. Listeners were asked to judge whether the stimulus sentences were uttered in isolation or in a paragraph, and if it was the latter, the position in the paragraph. Results showed that listeners were more likely to judge higher-pitched renditions of a sentence as paragraph-initial, indicating that pitch reset is a robust cue in discourse organization.

Ladd (1988) also showed that declination is an effective cue. He used four British English speakers to record sentences of an A and B but C (e.g., *ALlen is a STRONGer camPAIGNer, and RYAN has more POPular POLicies, but WARren has a LOT more MONey)*[4] or A but B and C format. Toplines were measured from the peak of the accented syllables. Results showed that the trend across the whole sentence is clearly downward, although downtrend of individual clauses also exists. Accented syllables after *but* is higher in F_0 than those after *and* when controlling for position. This is expected since disjunctures between *but*-conjoined clauses are higher in hierarchy than those between *and*-conjoined clauses. Ladd concluded that there is "partial reset" or "declination within declination" in sentences. Hierarchy affects the amount of reset.

Grosz and Hirschberg (1992) and Hirschberg and Nakatani (1996) tried to examine whether the relationship between structural hierarchy and declination and pitch

[4] The capitalized syllables are pitch-accented.

height holds using a read English corpus. They used discourse segment purpose (DSP), a discourse organizing unit proposed by Grosz and Sidner (1986), for segmenting discourse. A DSP is defined as an intention that the speaker tries to convey to the hearer. Results showed that DSP-initial phrases are characterized by larger pitch range and higher maximum and mean F_0 as compared to other utterance-initial phrases. On the other hand, DSP-medial and DSP-final phrases are lower in maximum and mean F_0.

Swerts and Geluykens (1994) showed that pitch height and declination are also indicative of topical boundaries in Dutch. Instructional monologs of three subjects from Terken (1984) were used and topic structures were segmented based on the instruction content. Acoustic analyses showed that F_0 peaks on topic introduction NPs are the highest. Relative heights of F_0 peaks seem to serve a dual function. On the one hand, they demarcate topical units; on the other hand, they highlight new topics by giving them the highest F_0 peak within a topical unit. There is also a global declination of F_0, which begins relatively high at the beginning of a unit, and it then slowly decreases over the course of the instruction.

Silent pause is another cue that has been proposed by a number of studies. For example, Lehiste and Wang (1977) and Lehiste (1979a) showed that listeners are more likely to judge a sentence as positioned at the end of a paragraph when it is followed by a long silent pause. Subjects were asked to mark sentence and paragraph boundaries by pressing a button when listening to normal and inverted (Lehiste & Wang, 1977) or low-pass filtered spontaneous speech (Lehiste, 1979a), where only some suprasegmental information was preserved. Results showed that the percept of a paragraph boundary is more likely to be indicated by longer pause as compared to sentential boundaries.

Using a larger but still read English corpus, Grosz and Hirschberg (1992) and Hirschberg and Nakatani (1996) again found there to be a correlation between duration of silent pauses and discourse hierarchy. They showed that DSP-initial phrases are preceded by longer pauses and followed by shorter pauses as compared to other utterance-initial phrases. On the other hand, DSP-final phrases have shorter preceding pauses and longer subsequent pauses compared to other utterance-final phrases. DSP-medial and DSP-final phrases are distinguished by the duration of the following pause. DSP-final phrases have longer pauses than DSP-medials. Traum and Heeman (1997) studied a spontaneous English dialog corpus and found similar results. Silence plays a role in relatedness when boundary tone is absent. A higher percentage of long silence (i.e., > 500 ms) follows an utterance that was related to the immediate previous utterance by another speaker as compared to one that is unrelated or related to not the immediate previous utterances.

Swerts and Geluykens (1994) also found silent pause to be a salient cue in Dutch. In the spontaneous instructional monologues examined, silent pauses are present at all transitions between topics, and also after the topic-introducing phrase or clause, the former being longer than the latter. Swerts and Geluykens suggested that pauses at the latter location reflect an interactive dimension, giving listeners the chance to process the new referent, but also giving them the opportunity to intervene if necessary, while pauses at the former location are of a demarcation function. In a subsequent perception experiment, they transformed the stimuli into four conditions: (1) no change; (2) pause duration was changed to a constant average pause; (3) melody was made to 200 Hz; and (4) pause and melody were both changed. For all four conditions, the signal was band-pass filtered (260–310Hz). Subjects were asked to determine when there was a major boundary. Results showed that both melody and pause duration contribute significantly to the perception of discourse structure in an additive manner. Locations where the two coincide elicited more responses. However, melody sometimes only contributes when pause variation is absent. Most of the responses were made before the beginning of the next topic. Swerts and Geluykens thus concluded that larger-scale information units are phonetically encoded in pausal and melodic properties, the former being more important that the latter. Finality cues and the subsequent long pauses are more important than initial cues, which are somewhat redundant and merely serve as confirmations that a new unit has started.

In addition to silent pause, Swerts (1997, 1998) also found that filled pauses (FPs), *um* and *uh* in particular, are robust indicators of discourse boundaries in spontaneous Dutch monologs. Discourse structure is determined by asking subjects to mark the "paragraph boundaries" when listening to and reading the utterances. Strong discourse boundaries are defined by when 75% of the subjects marked a boundary. Results showed that a majority of weak discourse boundaries have no FP (60%) while a majority of strong discourse boundaries have initial FPs (68%). Also, *uh* is more likely to appear in phrase-medial positions (85%) while *um* is more likely to appear in the phrase-initial position (79%). This is consistent with the view that *um* tends to signal planning of larger units while *uh* may be more likely for local lexical decision (Shriberg, 1994). Also, FPs at initial position tend to occur with flanking silent pauses (56%) while those in medial position tend to have no silent pauses (34%) or silent pauses afterwards (49%) and are thus integrated better into surrounding words. In terms of acoustic measures, FPs in initial positions have higher pitch and longer duration than those in medial positions.

There is also a general trend that *um* has higher pitch and longer duration than *uh* given the same position. This is in congruence with what Lehiste (1975a) has found in that higher-pitched elements are likely to be considered as paragraph-initial. Swerts thus suggested that FPs can act as a way of topic organization.

In comparison to pause, fewer studies have looked at the correlation between degree of final lengthening and discourse hierarchy. However, Lehiste and Wang (1977) and Lehiste (1979a) claimed that the two are positively correlated. Listeners are more likely to signal a paragraph boundary when the degree of preboundary lengthening is amplified.

Fon and Johnson (2000) used SOI as a combined measure for final lengthening and silent pause and found that the degree of final lengthening of SOI is proportional to the level of structural hierarchy in Guoyu spontaneous speech. Lengthening at the discourse level (using DSP as a unit) tends to be longer than lengthening at the sentential level. In other words, there is hierarchical encoding regarding sentential and discourse boundaries. Ladd (1988) used a similar duration measure termed "boundary duration" to look at English hierarchical encoding. This is defined as the interval between the onset of the last stressed syllable preceding the boundary and the onset of phonation of the first syllable after the boundary. In other words, this includes the duration of the final stressed syllable, the following unstressed syllables if there are any, and the following pause. Using read speech, Ladd showed that boundary duration is a good indicator of hierarchical organization of clauses. Clausal boundaries of higher disjunctures have longer boundary duration than those of lower disjunctures.

Speech rate is a derived measure from duration, and a couple of studies have proposed that it can somehow also reveal discourse structure. Hirschberg and Nakatani (1996) showed that in read English monologs, DSP-medial and DSP-final phrases differ in speaking rate in that DSP-medial phrases have slower speaking rates while DSP-final phrases are spoken faster. Koopmans-van Beinum and Donzel (1996) also showed that speech rate variation is indicative of information flow in Dutch spontaneous discourse. Pause-delimited units conveying new information are usually short and uttered at a slow rate while those conveying old information are usually long and spoken at a fast rate.

There is not much mentioning in the literature regarding amplitude patterning and discourse hierarchy. Hirschberg and Nakatani (1996) is probably one of the very few. They found that DSP-initial phrases are significantly higher in maximum and mean RMS amplitude as compared to DSP-medial and DSP-final phrases, indicating that amplitude can also serve as a possible cue for discourse organization.

Lehiste and Wang (1977) and Lehiste (1979a) proposed that laryngealization is also a cue that listeners use to induce a paragraph boundary percept as opposed to a sentential boundary. Kreiman (1982) agreed with this viewpoint, although she found that laryngealization sometimes also exists at sentential boundaries. Kreiman thus claimed that listeners are likely to posit a structural boundary of some kind if there is a change in voice quality from modal to creak.

All of the cues discussed above reveal at least some facet of reality concerning discourse hierarchy. However, the fact that most hierarchy cues distinguish themselves from boundary cues only via quantity, not quality, and that there is no absolute value set

to the magnitude of quantity makes one wonder whether and how listeners actually make use of these proposed cues in discourse processing.

Kreiman (1982) tried to answer this question by examining sentence and paragraph boundary cues in natural English conversation (i.e., a simulated telephone conversation). A paragraph boundary is defined as the end of a fairly complete unit of speech or the beginning of a new thought, regardless of speaker turns. Spontaneous dialogues were recorded and excerpts of 6 min were extracted from a 45-min conversation. The signal was first 200Hz low-pass filtered and then combined with a spectral inverted version. 8 subjects heard both the original and the transformed versions and were asked to indicate sentential and paragraph boundaries.

Results showed that the location of both sentential and discourse boundary responses is predicted by presence of a pause, laryngealization (creaky voice), pre-boundary lengthening, and/or a non-level intonation contour. However, nearly all paragraph responses in both normal and modified conditions came after the onset of the next utterance. Only in 13% of the normal condition and 22% of the inverted condition did subjects respond before. Thus, Kreiman claimed that paragraphs are not marked just by terminal contours, but are also cued by characteristics of the end of one utterance and the beginning of the next. Subjects are essentially comparing two blocks of speech. If the differences between them are big enough, a paragraph boundary is then assigned.

This is different from what Swerts and Geluykens (1994) have proposed for Dutch. Subjects in their study tended to respond to paragraph boundaries before the new unit appears. One reason for this discrepancy might be because that although both studies used spontaneous speech, the corpus used in Swerts and Geluykens is in fact more constrained than that used in Kreiman. The former is instructional in nature while the latter is more like a free telephone conversation. Thus, the cues pertaining to hierarchy may not be as obvious as those in more restricted monologs.

2.3 Summery

In this chapter, many cues in the prosodic and acoustic-phonetic domains pertaining to boundary and hierarchy information studies are discussed. However, it seems that not every cue is of the same importance. Prosodic cues, silent pause duration, and final lengthening are robust indicators for clausal and sentential boundaries. On the other hand, pause duration and pitch reset are strong indicators for hierarchy. Interested in both types of cues, this study intends to examine mainly prosodic boundary breaks, declination and pitch reset, and the final lengthening effect. In terms of final lengthening, although many units have been proposed to show this effect (e.g., foot in Lehiste (1975b), rhyme in Campbell and Isard (1991) and Oller (1973), syllable in Shen (1992), SOI in Fon & Johnson (2000), etc.), syllable and its derived measure SOI are used. This is in a way to accommodate the cross-linguistic nature of this study, since not every language has a canonical shape for foot (as the trochaic foot in English, Cutler & Carter, 1987), but all languages share the concept of syllable.

CHAPTER 3

METHODS

This chapter describes the methods used in recording and analyzing the cross-linguistic corpus for this dissertation project. Prosodic and acoustic-phonetic cue packages at sentential and discourse boundaries are examined to see if consistent patterns show up for the boundary effect, the hierarchy effect, and the language effect.

3.1 Subjects

Subjects from three languages—English, Japanese, and Mandarin, and two dialects of Mandarin—Guoyu and Putonghua, were recruited. All speakers were recruited from Columbus, Ohio. Japanese and Mandarin native speakers were recruited through student associations and personal connections. In order to attain homogeneity, only native speakers of Central Ohio English from Columbus and neighboring counties (Flanigan & Norris, 2000), Tokyo Japanese from Tokyo and three neighboring prefectures (i.e., Chiba, Saitama, and Kanagawa), Taipei Guoyu from Taipei and Taipei County, and Beijing Putonghua from Beijing area were included. Subjects were either born and raised in the language area or moved there before age three, and had no exposure to other languages before that. It was very difficult to find pure monolingual speakers for all three languages. English speakers from the local community often took classes in foreign languages during their high school and college years to fulfill the second language requirement. Of the eight English subjects recruited, only two claimed that they do not know any language other than English. For Mandarin and Japanese speakers, it was even more unlikely to record monolingual speakers, not only because they were recruited from the local community, but all speakers in the two language groups were required to take English as their second language in elementary or high school. However, Japanese and Mandarin speakers that lived in the States for more than three years, and had not used their native languages as the dominant everyday language since moving to the U.S. were excluded. Four female and four male subjects were recruited for each language/dialect group. Thus, there are 4 (subjects) × 2 (genders) × 4 (languages/dialects) = 32 subjects in total. Appendix A gives the detailed demographics of the subjects recruited.

3.2 Stimuli

The Pear Story (Chafe, 1980), a short film without spoken language, was played to the subjects in order to elicit speech. There are two reasons for choosing a film. First of all, a film does not provide an explicit structure, to which a narration has to adhere. Instead, viewers have to infer and reconstruct the structure from the film based on their own understanding and interpretation. Fon and Johnson (2000) used two four-frame comic strips to elicit speech. Although the material was easy to manipulate, the comic frames seemed to provide subjects with a priori boundaries in their narration. Since this study is also interested in how speakers structure their discourse (instead of how subjects react to a structured discourse), a film was used instead to circumvent this problem. Also, a cross-linguistic study requires relatively cultural-free material for elicitation. Although the comic strips used in Fon and Johnson were a successful tool for eliciting speech from Guoyu speakers, a preliminary survey of English speakers showed that most had some trouble understanding the message conveyed. Since *The Pear Story* was deliberately filmed to provide a relatively cultural-free story for discourse analysis studies, it is thought to be a good candidate for this experiment (e.g., English: Kärkkäinen, 1996; Mandarin: Erbaugh, 1990; Japanese: Clancy, 1980; Polish: Pakosz & Flaschner, 1988).

3.3 Equipments

Recordings were done with a SHURE SM10A head-mounted microphone connected to a SONY DAT DTC-790 recorder through a Symetrix SX202 Dual Mic Preamp preamplifier, and using Maxell R-64DA DAT tapes. A D-to-D transfer was done and the sampling rate was set at 44100 Hz and was later downsampled to 22050 Hz for further analyses.

3.4 Procedure

Recordings were made in the Phonetics Laboratory in the Department of Linguistics, The Ohio State University. Subjects were tested individually in a quiet room. To avoid any accommodation effect that subjects might have, the author asked an

experimenter who is also a native or a near-native speaker of the dialect of one of the three languages to conduct the experiment for each language group.[1] Subjects were first shown the short film and were then asked to describe it afterwards as if talking to a friend. Most of the subjects had no trouble with the instruction. Some of the subjects told the stories more than once due to various reasons. Subject HPH recorded three times because he did not like the first try due to its informal register. The second try was self-interrupted and the third try was rather formal. The first recording was used for further analyses. Subject WCF requested to do a practice storytelling first before the actual recording. Subject FBS was interrupted during the first try because he was using a fairly literary and formal register. He then rerecorded twice, for he was not satisfied with the second recording. The third recording, which was the more informal of the two, was used for analysis. The experiment took about 20–30 min, and each subject was paid $5 for his participation. Appendix B shows the duration of the recording of each subject.

3.5 Transcription and Prosodic Labeling

Recordings were first orthographically-transcribed by (near-)native speakers.[2] Afterwards, they were prosodically partially transcribed by experienced transcribers of

[1] The Guoyu and Japanese groups were recorded by native speakers of the two languages. The author, who was born and for the most part of her life, raised in Taipei, recorded the Guoyu group. The Japanese experimenter, Kiyoko Yoneyama, was born and raised in Tokyo. English and Putonghua were recorded by near-native speakers of the dialects recorded. There were two English experimenters: Keith Johnson is an English speaker from Oklahoma, and Terah Schamberg is from Orwell, a northwestern town in Ohio. The Putonghua experimenter, Tsan Huang, was originally from Rugao, a city in Jiangsu Province in the east coast of China. At age 7, she started to learn Putonghua in school, the common language for pedagogy. She went to Beijing for college when she was 18. After graduating from college, she taught at a university there for 8 years before coming to the United States.

[2] Both varieties of Mandarin were transcribed by the author and Putonghua was double-checked by the Putonghua experimenter Tsan Huang. English transcription was done by Terah Schamberg and

the Tone and Break Indices (ToBI) systems of respective languages.[3] English ToBI (E-ToBI) transcription was based on Beckman and Ayers Elam (1997), Japanese ToBI (J-ToBI) was based on Venditti (1997), and Mandarin ToBI (M-ToBI) was based on Peng et al. (2000) and Tseng and Chou (1999). ToBIs of different languages have slightly different requirements for labeling tiers. E-ToBI (Beckman & Ayers Elam, 1997) consists of four tiers: a tone tier, an orthographic tier, a break-index tier, and a miscellaneous tier; J-ToBI (Venditti, 1997) has five tiers: word tier, tone tier, break index tier, finality tier, and miscellaneous tier. The finality tier is sometimes combined with the break index tier at some sites; and M-ToBI (Peng et al., 2000) requires seven tiers: romanization tier, syllable tier, stress tier, sandhi tier, tone tier, break index tier, and code tier. However, all ToBI systems invite free addition of site-specific tiers.

Since ToBI labeling is very time-consuming, and since this study is mainly focused on prosodic boundaries (i.e., break indices) rather than tonal prominence, some of the tiers are left out to make the project more doable. The tiers that are left out include the finality tier in J-ToBI, which is combined with the break index tier, and the romanization tier, the stress tier, the sandhi tier, and the code tier in M-ToBI.

There are differences among the transcription tiers in the three ToBI systems. E-ToBI and J-ToBI adopt a word-based unit for transcription (i.e., the orthographic tier in E-ToBI and the word tier in J-ToBI), while M-ToBI uses a syllable-based unit for transcription (i.e., the syllable tier). To accommodate the three, both word and syllable tiers are included in the generic labeling tiers in this study. In other words, E-ToBI and J-ToBI transcriptions are augmented with a syllable tier and the M-ToBI transcriptions are augmented with a word tier. Perhaps due to the syllable-per-character writing system in Mandarin, the definition of word is a matter of debate. The definition provided by Li and Thompson (1981) is adopted here as a convenient solution. Since the main interest of this study concerns the mapping between structural and prosodic hierarchies, two tiers regarding structural boundaries, syntax and discourse tiers, were also added to the system. Another tier regarding pitch trend was also added to facilitate analyses.

In total, there are eight tiers in the generic ToBI labeling system adopted in this study. As shown in Figure 3.1, the topmost tier is the tone tier, which is present in all three ToBI systems. It labels accents and boundary tones of utterances. The next two tiers are the word and syllable tiers. The pitch trend tier labels the high tonal targets around the boundary region and will be described in Section 3.6. The discourse tier labels discourse boundaries and will be explained in Section 3.7. The syntax tier labels syntactic relations between two adjacent clauses, and does not concern us here. The break index tier is also present in all three ToBI systems. It labels the sizes of prosodic breaks. Finally, the miscellaneous tier is also included. This tier is present in both E-ToBI and J-ToBI and labels events that do not belong to the other tiers, such as silent pause and laughter.

3.6 Acoustic Measurements

Two acoustic measurements were taken—F_0 peaks and duration. The first measurement type targets tone target locations within tone-bearing syllables whereas the

double-checked by the author. Japanese transcription was transcribed by Aya Shirai, a native speaker of Tokyo Japanese and double checked by Kiyoko Yoneyama and the author.

[3] English ToBI transcription was done by Craig Hilts and Japanese ToBI transcription was done by Jennifer J. Venditti. Due to immaturity of Mandarin ToBI, transcription of the two dialects of Mandarin was transcribed by and the results discussed between Tsan Huang and the author.

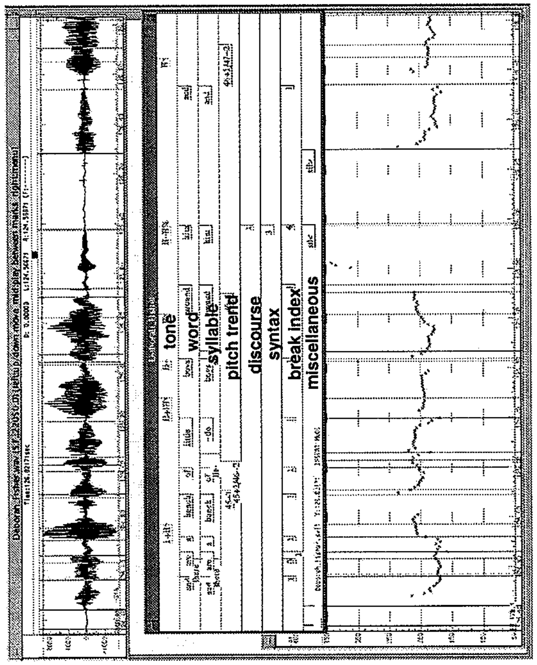

Figure 3.1: An illustration of the modifications made on the layout, which is based on a generic ToBI labeling system. There are eight tiers in total. Please see the text for description.

second measurement type targets intervals corresponding to syllables and pauses. Syllable is a common prosodic unit across languages, although syllable structures (and hence typical syllable lengths) are different. English has the most complex syllable structure and Japanese the simplest. The most complicated syllable structure in Japanese and Mandarin is perhaps CGVN, which is nowhere as loaded as the CCCVCCC structure in English (e.g., *strengths*). Syllable breaks in English were determined by *Merriam-Webster Online Collegiate Dictionary* (2002). In Japanese, geminate consonants were grouped with the following syllable so that words such as *kitto* 'definitely' were segmented as *ki.tto*.

Silent pause duration was measured whenever there was a perception of such by the transcribers instead of setting an arbitrary cutoff point. This is because speech rate fluctuates to a great extent in spontaneous speech. It thus makes more sense to judge silent pause in context and use trained ears as a way of normalization to compensate for what a fixed cutoff point could not achieve. Both syllable and pause were measured in terms of milliseconds and as a result, two duration measures are taken. One is the syllable duration, and the other is the syllable onset interval (SOI), which is the interval between one syllable onset to the next syllable onset. This includes essentially the duration of a syllable and whatever pause that follows it (Fon & Johnson, 2000).

Duration measures are based on the syllable. Since what is of interest in this study is discourse and sentential boundary cues, it would be reasonable to focus measurements and analyses on the boundary region. The syllable that is immediately before the boundary should reflect the most influence of the boundary in the duration domain. It therefore is designated the boundary syllable symbolized as P_0 (standing for Position

Zero) in this study. To obtain reference values for how the boundary effect might influence the measures, duration of two syllables before and after P_0 were also calculated. The two syllables before P_0 will be referred to as P_{-2} (i.e., two syllables before P_0) and P_{-1} (i.e., one syllable before P_0), and the two syllables after will be referred to as P_1 (i.e., one syllable after P_0) and P_2 (two syllables after P_0). A schematic illustration of the five positions chosen for further analyses is shown in Figure 3.2.

By taking two reference syllables before and two after P_0, one could also examine whether the boundary effect is focused on only the boundary syllable or is more spread out. Thus, for both duration measures, syllable duration and SOI, five measurements will be taken around the boundary region. In other words, in the syllable duration measure, the five values refer to the duration of the five syllables around the boundary region, the duration of possible postsyllabic pauses being ignored. Similarly, in the SOI measure, the five values refer to the duration of the five SOIs around the boundary region.

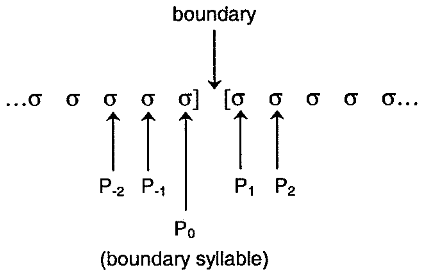

Figure 3.2: A schematic illustration of how the five syllables around the boundary region were designated for duration analyses.

F_0 peaks were obtained from syllables bearing high tones. Only high tones are used as targets for pitch trend analyses because studies have shown that the lower end of the range does not change much for the kinds of pitch range manipulations typical of monolog narratives (Ladd, 1988). High tones are defined language-dependently. In English, high tones are pitch-accents that contain an H defined in the ToBI prosodic system promoted by Beckman and Ayers Elam (1997). This includes H*, L+H*, L*+H, H+!H*, !H*, L+!H*, and L*+!H. In Guoyu and Putonghua, high tones are defined as the high tonal targets of Tone 1 and Tone 4. Since Putonghua makes a distinction between stressed and unstressed syllables, care was taken so that only high tonal targets of stressed Tone 1 and Tone 4 were chosen to exclude occasions of target undershoot.[4] For Japanese, high tonal targets of pitch-accented syllables, which are always high, were chosen. F_0 extraction was made possible through Perl scripts, and F_0 maximum of each high-toned syllable was extracted.

As with the duration measure, the patterning of the pitch trends at the boundary region is of the most interest. However, unlike syllables, high tones are not as densely distributed. Therefore, as shown in Figure 3.3, only four points, P_{-2}, P_{-1}, P_1, and P_2, were chosen. Figure 3.4 shows an actual example of how the high tonal points were selected. Notice that for each point, the F_0 maxima of the high tonal targets were selected.

[4] Both Tsan Huang and the author noted that sometimes there is an intermediate level of stress in Putonghua. That is, tonal contours are realized fully but they are somehow perceptually not as prominent as other also realized tonal contours. Those syllables were not chosen for pitch trend analyses.

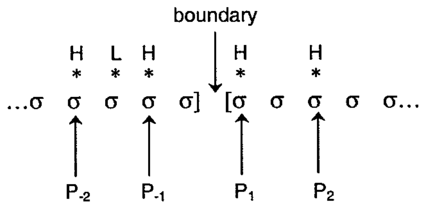

Figure 3.3: A schematic illustration of how the four high tones around the boundary region were chosen. Asterisks indicate pitch-accented syllables.

3.7 Discourse Labeling

Since sentence is a difficult unit to define and identify in spontaneous speech, clause was used as a basic unit instead. A clause was defined by having at least one main verb. This definition suffices with the English and Japanese corpora.[5] However, Since Mandarin has no morphological markings differentiating serial verbs from sequences of main verbs, it is sometimes difficult to segment utterances into clauses. In this study, the guidelines laid out by Li and Thompson (1981) were adopted for Mandarin clause segmentation. For all 32 subjects in the database, utterances were segmented into clauses, and for every pair of adjacent clauses, their relationship was determined by using the guidelines outlined in Grosz and Sidner's (1986) computational psycholinguistic

[5] Criteria for Japanese clause segmentation were determined in consultation with Dr. Mineharu Nakayama and Dr. Akihiro Kano.

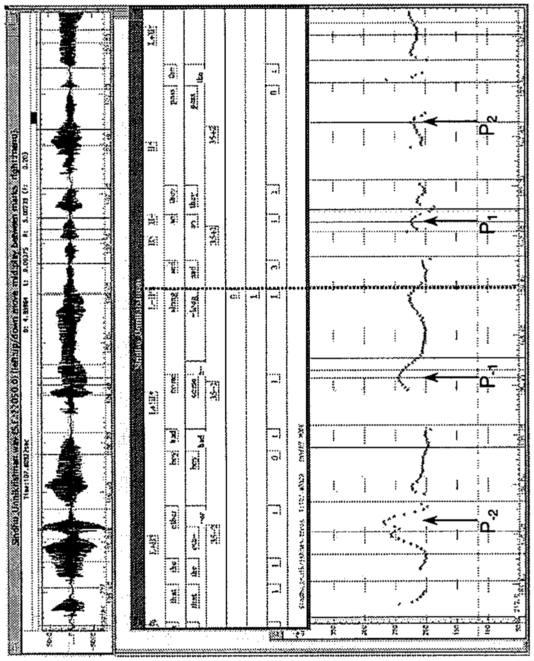

Figure 3.4: An actual example of how the four high tones around the boundary region were chosen. The dotted vertical line indicates the structural boundary.

discourse framework. Three levels of disjunctures—DSP0, DSP1, and DSP2—were identified. Definitions of each level and examples are given below.

DSP0 refers to the lowest possible degree of disjuncture between two adjacent clauses in a discourse. Namely, no discourse level disjuncture exists. The only boundary is the syntactically defined clausal boundary. DSP0 is relevant because it represents a potential discourse disjuncture. Examples from the three languages are as shown in (1)–(3). For non-English data, the first line is the romanization of the utterance,[6] the second line is a word-by-word literal translation, and the third is a rough English translation of the utterance. The IDs in parentheses indicate from which narrative the utterances have been taken.

(1) English:
[*this little boy comes along on his bicycle*]$_{DSP0}$ [*and uh notices the basket of pears*]
(SU)

(2) Mandarin:
[*ta you kan-le-kan shu-shang nei-ge guonong*]$_{DSP0}$ [*kan ta mei zhuyi*]
[he again watched-a -little tree-on that-CL[7] fruit-farmer]$_{DSP0}$ [see him not notice]
[he watched (the) fruit farmer on the tree for awhile]$_{DSP0}$ [found him not noticing]
(CY)

(3) Japanese:
[*sono hito ga ki ni nobotte*]$_{DSP0}$ [*Etto nashi o moideita to omouN desu*]
[that person nom. tree to climb-and]$_{DSP0}$ [uh pear acc. picked comp. think BE]
[that person climbed up (the) tree and]$_{DSP0}$ [I think (he) picked pears]
(FY)

[6] In this study, Mandarin romanization follows the Pinyin convention and Japanese romanization of the corpus follows the convention laid out in Venditti (1997). Long vowels are indicated by capitalization.
[7] CL: classifier.

As shown in (1)–(3), at the DSP0 level, the two clauses share the same communicative intentions. The tight linkage is also evidenced syntactically by frequent zero anaphora and the fluent conjunction *and* (Bestgen, 1998).

DSP1 refers to the next higher level of disjuncture. It designates a boundary break between two adjacent clauses that is interrupted by not only a clausal boundary, but also a small discourse boundary. The two clauses are used to communicate two different discourse purposes. However, these two purposes are related to each other. Examples from the three languages are given below in (4)–(6).

(4) English:
[*and then a little boy came by on a bicycle*]$_{DSP1}$ [*and he stole a bushel of pears on this too big of a bicycle*]
(MD)

(5) Mandarin:
[*jiu ba ta zai-zou-le*]$_{DSP1}$ [*zhe-ge jiao touqie ba*]
[then acc. it carry-away-ed]$_{DSP1}$ [this-CL call stealing part.][8]
[(he) then carried it away]$_{DSP1}$ [this is called stealing, no?]
(CHL)

(6) Japanese:
[*de sorede Etto aoi nasi o totteite*]$_{DSP1}$ [*de aru tEdo tamatte*]
[then thus uh green pear acc. picked-and]$_{DSP1}$ [then some degree stack (v.)-and]
[(he) then uh picked green pears and]$_{DSP1}$ [then (the pears) was stacked to some degree and]
(YE)

[8] Part: particle.

As shown in Examples (4)–(6), adjacent clauses with a DSP1 level disjuncture in-between refer to clauses that belong to different discourse intentions. This is often evidenced by a subject change.

Finally, DSP2 refers to the highest level of disjuncture in this study. Similar to DSP1, it designates a boundary break between two adjacent clauses that is interrupted by both a clausal boundary and a discourse boundary. In addition, the two clauses are devoted to two unrelated discourse purposes. Examples of the three languages are given in (7)–(9).

(7) English:
[goat man with his goat passes]$_{DSP2}$ [he comes down][9]
(SU)

(8) Mandarin:
[*ta ye bu zaiyi*]$_{DSP2}$ [*zhe-shihou ne you you yi-ge xiaohair*]
[he also not care]$_{DSP2}$ [this-time part. again have one-CL child]
[he doesn't care]$_{DSP2}$ [then there is another child]
(LX)

(9) Japanese:
[*E kyO E bideo o mimasite*]$_{DSP2}$ [*hazimari ga tIsana kodomo ga E deNsya ni notte*]
[uh today uh video acc. watched-and] [beginning nom. small child nom. uh bike on ride-and]
[uh today uh (I) watched (a) video and]$_{DSP2}$ [in the beginning a child rode a bike and]
(NT)

[9] This *he* refers to the pear farmer and is not coreferenced to the *goat man*.

As shown in (7)–(9), DSP2 level disjunctures refer to the situation when two adjacent clauses are serving two unrelated discourse purposes. This is often indicated syntactically by a full anaphora and different subjects.

In summary, three levels of discourse disjuncture are recognized—no discourse disjuncture (DSP0), low discourse disjuncture (DSP1), and high discourse disjuncture (DSP2). The following Chapters 4–6 are devoted to the results concerning the various discourse disjuncture cues in the prosodic and acoustic-phonetic domains.

CHAPTER 4

RESULTS: PROSODIC BOUNDARY BREAKS

Although prosody has a hierarchical organization in its own right (Beckman, 1996; Selkirk, 1984, 1986), there is also evidence that prosodic and clausal/sentential boundaries coincide to a large extent (Price, Ostendorf, Shattuck-Hufnagel, & Fong, 1991; Speer, Kjelgaard, & Dobroth, 1996). It is therefore not illogical to hypothesize that prosodic hierarchy would somehow reflect discourse hierarchy, at least in some instances (Swerts, 1997; Terken, 1984; Traum & Heeman, 1997). In this chapter, the relationship between prosodic breaks and structural boundary and hierarchy is examined.

4.1 ToBI Comparisons

Although there are many prosodic frameworks to choose from, not many have proven themselves to be able to break away from language-specificity. Since it is essential in a cross-linguistic study to use equivalent frameworks for each of the languages, a prosodic labeling system that has its root in language-universals but also takes into account cross-linguistic prosodic differences is necessary. The ToBI (Tone and

Break Indices) labeling system fits the purpose of this study because it employs the generic principles that underlie language universality while at the same time accommodates language specificity to a large extent. Moreover, there are at least already some studies in the literature concerning the labeling systems of the three languages in question so that one does not have to start from scratch (English: Beckman & Ayers Elam, 1997; Mandarin: Peng et al., 2000, Tseng & Chou, 1999; Japanese: Venditti, 1997). There are two parts to the ToBI labeling system—the tones and the break indices (BI). The current chapter focuses on the latter, since it is considered more relevant to the research questions of this study, boundary and boundary hierarchy.

Because the ToBI system accommodates individual language needs, the basic unit for assigning break indices is different for the four languages. For English and Japanese, the basic unit is the word. That is, every word boundary is assigned a break index. On the other hand, for Mandarin, break indices are assigned to every syllable boundary instead. Since this study is interested in both language universals and language specifics, the smaller of the unit, the syllable, is adopted as the basic unit for study in this chapter. This decision is based on two reasons. First of all, syllable is also used as the basic unit in duration analyses. Using syllable as the basic unit here would make the analyses more comparable. Also, by adopting the smaller of the units, one could still look at the phenomenon that is related to the bigger of the units, which is not possible if it were the other way around.

Table 4.1 shows the distribution of break indices of the four languages using syllable as the unit. In the spirit of accommodating language-specificities, the break index systems for the four languages are somewhat different. In English ToBI (hereafter

	English	Guoyu	Putonghua	Japanese
BI0	498	805	478	0
BI1	1836	4716	2727	1583
BI2	31	691	1933	238
BI3	220	569	559	1081
BI4	643	297	168	98
BI5	---	104	42	---
BI0/2	0	11	0	0
BI0/4	1	1	0	0
BI0p	---	---	---	57
BI1p	29	---	---	2
BI2p	71	---	---	4
BI3p	43	---	---	4
BI2-	0	---	---	85
BI3-	0	---	---	107
BI2m	---	---	---	29
BI3m	---	---	---	44
No BI	646	---	---	3372
Total	4018	7194	5907	6704

Table 4.1: Distribution of break indices (BIs) for the four languages using syllable as the unit. 'No BI' indicates there is no BI applicable to the syllable break due to an absence of a word boundary. '—' indicates that the category is not applicable in that particular ToBI system.

BI	Definition
BI0	for cases of clear phonetic marks of clitic groups
BI1	for most phrase-medial word boundaries
BI2	for (1) a strong disjuncture marked by a pause or virtual pause, but with no tonal marks; or (2) a disjuncture that is weaker than expected at what is tonally a clear intermediate or full intonation phrase boundary
BI3	for ip boundary
BI4	for IP boundary
BI1-	for uncertainty between BI0 and BI1
BI2-	for uncertainty between BI1 and BI2
BI3-	for uncertainty between BI2 and BI3
BI1p	for abrupt cutoff
BI2p	for prolongation
BI3p	for hesitation after the onset of the tonal marks for an ip

Table 4.2: Definitions of BIs for E-ToBI. Adapted from Beckman & Ayers Elam (1997). IP: intonation phrase; ip: intermediate phrase.

E-ToBI), there are five basic levels of disjuncture. BI1 is the "default" word boundary break. BI0 is for clitic groups such as *got you* → *gotcha*. BI2 is for mismatches between tone and disjuncture. BI3 and BI4 are for intermediate (ip) and intonation phrase boundaries (IP), respectively. Uncertainties are marked by minus signs and disfluencies by *p*'s. Table 4.2 shows the definitions of break indices in E-ToBI (Beckman & Ayers Elam, 1997).

Table 4.3 shows the definitions of BI of the Japanese ToBI (hereafter J-ToBI) system (Venditti, 1997). Like the E-ToBI, it has five basic BI levels. BI1 is again the "default" word boundary. BI0 is for certain lenition processes such as contractions, for example, /yatte+simau/ → [yattyau] 'do completely'. BI2 and BI3 represent accentual (AP) and intonational phrase (IP) disjunctures, respectively. BI4 is for intonational phrase disjunctures with a strong sense of finality.[1] Like the E-ToBI, uncertainties and disfluencies are marked with minus signs and *p*'s, respectively. In addition, tonal and disjuncture mismatches are marked with *m*'s.

BI	Definition
BI0	for strong cohesion, typical of fast speech or AP-medial lenition processes
BI1	for no higher-level juncture, typical of the majority of AP-medial word boundaries
BI2	for medium degree of disjuncture, typically corresponds to the tonally-defined AP
BI3	for strong degree of disjuncture, typically corresponds to the tonally-defined IP
BI4	for IP boundaries with a stronger sense of finality/completeness
BI1-	for uncertainty between BI0 and BI1
BI2-	for uncertainty between BI1 and BI2
BI3-	for uncertainty between BI2 and BI3
BI0p	for word-internal disfluency
BI1p	for AP-internal disfluency
BI2p	for AP-final but IP-internal disfluency
BI3p	for IP-final disfluency
BI2m	for medium disjuncture with IP boundary tone
BI3m	for large disjuncture with no sign of range reset

Table 4.3: Definitions of BIs for J-ToBI. Adapted from Venditti (1997). AP: accentual phrase; IP: intonation phrase.

[1] BI4 is sometimes labeled as BI3 and finality is labeled in a separate tier. In this chapter, BI4 is used in the J-ToBI to make the system more comparable to the Mandarin ToBI system.

Of the three ToBI systems, Mandarin is the least developed. According to the Pan-Mandarin ToBI guideline (hereafter M-ToBI) in Peng et al. (2000) and the break index system developed by Tseng and Chou (1999), there are six levels of BIs (Table 4.4). Again, as in the E-ToBI and the J-ToBI, BI1 is the default boundary. However, unlike the two ToBIs, the unit of disjuncture labeling is the syllable instead of the word. This is due to a strong syllable-character correspondence in the language. BI0 is for reduced syllables, for example, *ni3-men0* → *nim3* 'you (pl.)'. BI2 and BI3 are used for phrase boundaries within a breath group. BI3 is distinguished from BI2 by a following pause. BI4 and BI5 indicate breath group boundaries, with BI5 characterized by a prolonged pause.

If one disregards categories concerning uncertainties, disfluencies, and mismatches, there are five basic BI levels in the E-ToBI and J-ToBI and six in the M-ToBI. As shown in Table 4.5, BI0 and BI1 are unanimous across all three labeling systems in that BI0 represents some kind of contraction or deletion at the word or syllable

BI	Definition
BI0	for reduced syllable boundary
BI1	for normal syllable boundary
BI2	for minor phrase boundary, not followed by a pause
BI3	for major phrase boundary, followed by a pause
BI4	for breath group boundary, including a pitch reset
BI5	for prosodic group boundary, followed by a prolonged pause

Table 4.4: Definitions of BIs for M-ToBI. Adapted from Peng et al. (2000) and Tseng and Chou (1999).

level and BI1 is used for the default syllable/word boundaries. BI2 in the E-ToBI is unique in that it represents a mismatch between tone and disjuncture. This is represented in the J-ToBI by the diacritic *m* and is not represented by the M-ToBI at all. According to Beckman and Pierrehumbert (1986), an AP is a level between a prosodic word and an intermediate phrase. However, E-ToBI only recognizes ip's (represented by BI3) while J-ToBI only recognizes APs (represented by BI2).

M-ToBI, on the other hand, recognizes two intermediate level phrase boundaries, BI2 and BI3. Although the M-ToBI manual does not define clearly what "minor" and "major" phrase boundaries are, the two seem to constitute a level that is intermediate,

E-ToBI	J-ToBI	M-ToBI	Definitions
BI0	BI0	BI0	for some kind of contraction/deletion
---	---	BI1	for "default" syllable boundaries
BI1	BI1	---	for "default" word boundaries
BI2	---	---	for mismatches between tone and disjuncture
---	BI2	BI2	for AP boundaries, minor phrase boundaries
BI3	---	BI3	for ip boundaries, major phrase boundaries
BI4	BI3	BI4	for IP boundaries
---	BI4	BI5	for IP boundaries with strong sense of finality

Table 4.5: A rough comparison of the labeling systems among the E-ToBI, J-ToBI, and M-ToBI. '---': nonapplicable.

with BI3 representing a slightly bigger disjuncture than BI2.[2] Therefore, BI2 in the M-ToBI is assigned to a level roughly equivalent to the BI2 in the J-ToBI (the AP boundary) since it is a disjuncture that is deemed to be slightly higher than a prosodic word but lower than an ip boundary, and BI3 in the M-ToBI is assigned to a level roughly equivalent to the BI3 in the E-ToBI (the ip boundary) since its disjuncture size is between a BI2 and an IP boundary. By putting BI2 in M-ToBI on the same level as BI2 in J-ToBI, one does not imply that a minor phrase boundary is an exact equivalent of an AP in J-ToBI, since whether Mandarin has an accent is still of debate. Different languages have different ways of organizing prosody, and they are often not exactly comparable. Table 4.5 is just a rough comparison showing the different levels of prosodic boundary breaks the three languages have and how they roughly compare to one another.

The highest level of disjuncture in the E-ToBI is BI4, which corresponds to an IP boundary. J-ToBI and M-ToBI have similar counterparts, BI3 and BI4 in the J-ToBI versus BI4 and BI5 in the M-ToBI. Although in the M-ToBI manual, BI4 is defined as a breath group boundary, it is deemed as a rough equivalent to an IP boundary because it requires a pitch reset. Finally, both the J-ToBI (BI4) and the M-ToBI (BI5) have an extra level of BI to indicate finality, which is missing in the E-ToBI convention. This is in fact due to different philosophies underlying the three systems. BIs up to the IP level encode prosodic breaks that are categorically defined. The kind of continuous variation that cues discourse structures directly and iconically is not tagged with the BI levels in the E-ToBI.

[2] For one of the M-ToBI labelers, Tsan Huang, BI2 in M-ToBI is the most difficult to decide and pinpoint. However, a close examination of the distribution of BI2's in Putonghua shows that it is not a level that is equivalent to a prosodic word. It is used as a label for groupings that is somewhat higher than a prosodic word.

In the J-ToBI and M-ToBI systems, however, this is not the case. A level that indicates discourse hierarchy is also labeled through BI4 in the J-ToBI and BI5 in the M-ToBI.

4.2 Data Selection Criteria

BIs at the boundary region with diacritics and units with more than one assigned BIs (see Table 4.1) were excluded from all further analyses since they imply uncertainty, disfluency, or mismatch. BI2 in E-ToBI is also excluded for the same reason. Appendix C shows the distribution of BI in each language and Appendix D shows the number of cases each subject contributed.

4.3 Predictions

Although not an exact match, prosodic boundaries are often thought to coincide with syntactic boundaries (Price, Ostendorf, Shattuck-Hufnagel, & Fong, 1991; Speer, Kjelgaard, & Dobroth, 1996). Therefore, a boundary effect is expected. This implies that BIs of higher values are more likely to occur at the structural boundary position than within a structural segment. Since where a BI is located determines the value it will assume, in the following analyses, this will be called the position effect. Also, one would predict that bigger structural boundaries have more higher BI levels. In other words, a hierarchy effect is also expected. With regards to the language effect, it is unclear how language might influence the patterning of BI distribution in addition to influences from different ToBI systems. However, such an effect would not be surprising.

4.4 Overall ANOVA

As shown in Table 4.5, the only language-universal BI level is that of an IP and above. Therefore, the percentage of IP (i.e., BI3 in the J-ToBI, BI4's in the E-ToBI, J-ToBI, and M-ToBI, and BI5 in the M-ToBI) was used as a dependent measure. The percentage values were arcsine transformed to attain a more normal distribution.

An overall Position (5) × Hierarchy (3) × Language (4) three-way mixed design ANOVA was performed to test the above predictions, the within-subject variable being position, and the between-subject variables being language and hierarchy.[3] There are four levels of language—English, Guoyu, Putonghua, and Japanese, three levels of hierarchy—DSP0, DSP1, and DSP2, and five levels of position, which are the five syllables selected from around the boundary region, three before the boundary, and two after. The dependent variable is the percentage of IP after arcsine transformation.

The ANOVA results are shown in Table 4.6. All of the effects are significant, providing a promising picture that all three factors are at work in determining the distribution of IP prosodic breaks at structural boundaries. In the following sections, 4.5 through 4.8, detailed descriptions of the post-hoc test results are given and the patterning of prosodic breaks of each language is described.

[3] Hierarchy is a between-subject variable due to the spontaneous corpus design and the selection criteria. To be a within-subject variable statistically, one has to have approximately equal number of cases from each category—DSP0, DSP1, and DSP2, which was not the case. Some subjects contributed more than others (see Appendix D). As a consequence, hierarchy was statistically a between-subject variable in this study, although almost all subjects had all three disjunctures in their narratives.

Source	*df*	*F*	$\hat{\eta}^2$	*p*
		Between subjects		
Language (L)	3	96.18**	.78	.00
Hierarchy (H)	2	43.82**	.51	.00
L × H	6	2.91*	.17	.01
Error	84	(0.02)		
		Within subjects		
Position (P)	1.80	842.25**	.91	.00
P × L	5.40	59.23**	.68	.00
P × H	3.60	27.61**	.40	.00
P × H × L	10.81	2.53**	.15	.01
Error	151.30	(0.04)		

Note. Values enclosed in parentheses represent mean square errors. Due to violation of sphericity assumption using Mauchly's W test ($p < .01$), the *df*'s in the within-subjects section were adjusted using Huynh-Feldt adjustments.

*$p < .05$. **$p < .01$.

Table 4.6: Overall ANOVA results for prosodic breaks.

4.5 English Results

There are five possible BI values an English syllable can assume here—BI0, BI1, BI3, BI4, or no BI. Figure 4.1 shows the distribution of prosodic breaks in English in terms of proportions. One sees that at each DSP level, BI1 is the most common break value, reflecting the large number of monosyllabic words in the English corpus. Also, there seems to be a complementary distribution between BI4, and BI0 and BI1 at the boundary region. BI4 is most likely to occur at P_0 while BI0 and BI1 are most likely to occur at other non-boundary positions. There are not many BI3s at the boundary region in general. Proportionately more boundary syllables take the break value of BI4 at higher discourse disjunctures. At P_1, there is also proportionately more BI3 and BI4 at the DSP2 level as compared to the DSP1 and DSP0 levels.

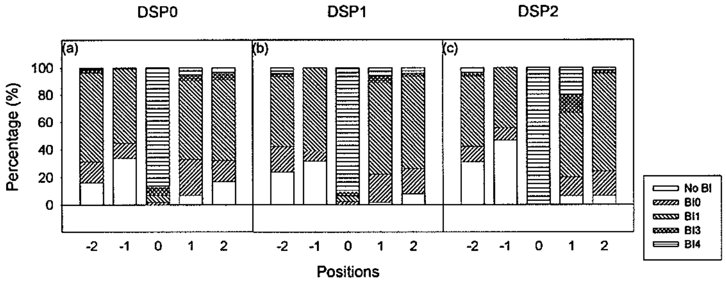

Figure 4.1: Bar graphs showing the percentages of BIs with regards to positions and discourse hierarchy in English. The *x*-axes represent position. '-1' and '-2' refer to one and two syllables before the boundary syllable, respectively, and '1' and '2' refer to one and two syllables after the boundary syllable, respectively. '0' refers to the boundary syllable. The *y*-axes represent percentages. 'No BI' represents syllable boundaries that do not coincide with word boundaries. (a) DSP0; (b) DSP1; (c) DSP2.

A post-hoc two-way Position (5) × Hierarchy (3) mixed design ANOVA on percentages of IP breaks was performed to confirm the above observations. As shown in Table 4.7, both the main effects and the interaction effect are significant. Horizontally across different positions, post-hoc analyses using Bonferroni's adjustment showed that IP breaks are more likely to occur at P_0 at all three levels ($p < .01$). Hierarchically across

Source	*df*	*F*	η^2	*p*
		Between subjects		
Hierarchy (H)	2	7.66**	.42	.00
Error	21	(0.03)		
		Within subjects		
Position (P)	1.70	241.95**	.92	.00
P × H	3.41	4.22**	.29	.01
Error	35.79	(0.07)		

Note. Values enclosed in parentheses represent mean square errors. Due to violation of sphericity assumption using Mauchly's W test ($p < .01$), the *df*'s in the within-subjects section were adjusted using Huynh-Feldt adjustments.
$*p < .05$. $**p < .01$.

Table 4.7: Post-hoc ANOVA results for proportions of IP breaks in English.

three levels, post-hoc analyses using the Tukey's-*b* test showed that the proportions of IP breaks at P_0 and P_1 reflect boundary strengths. For both positions, IP proportions at the DSP2 level are significantly higher those at DSP1 and DSP0 levels ($p < .05$).

In general, the statistical analyses confirmed the observations above. There is a robust position effect at P_0. In addition, the hierarchy effect shows a significant trend at P_0 and P_1.

4.6 Guoyu Results

There are six possible values for prosodic breaks to assume in Guoyu—BI0, BI1, BI2, BI3, BI4, and BI5. Figure 4.2 shows the distribution of prosodic breaks in Guoyu.

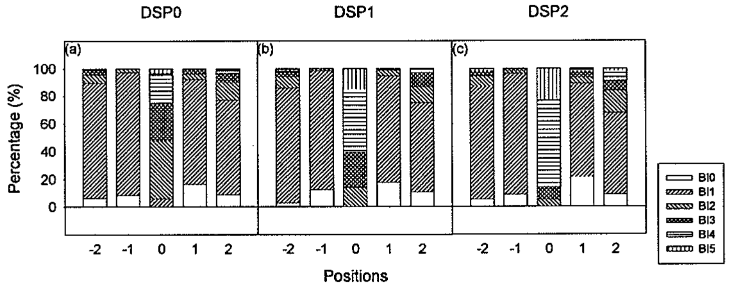

Figure 4.2: Bar graphs showing the percentages of BIs with regards to positions and discourse hierarchy in Guoyu. The layout of the graph is similar to that in Figure 4.1. (a) DSP0; (b) DSP1; (c) DSP2.

BI1 is the most common value for breaks. In addition, there seems to be a complementary distribution between BI3, BI4, and BI5, and BI0 and BI1 around the boundary location. The former group is most likely to occur at P_0 while the latter group is most likely to occur at other non-boundary positions. There are not many BI2s in general. The proportions of BI4 and BI5 at the boundary syllable increase as discourse disjuncture becomes bigger.

A post-hoc two-way Position (5) × Hierarchy (3) mixed design ANOVA on percentages of IP breaks (i.e., BI4 and BI5) was performed to confirm the above observations. Table 4.8 shows the results. Both of the main effects and the interaction effect are significant. Horizontally across different positions, post-hoc analyses using Bonferroni's adjustment showed that IP breaks are more likely to occur at P_0 at all three

Source	df	F	$\hat{\eta}^2$	p
		Between subjects		
Hierarchy (H)	2	26.83**	.72	.00
Error	21	(0.01)		
		Within subjects		
Position (P)	1.38	104.78**	.83	.00
P × H	2.76	17.42**	.62	.00
Error	28.94	(0.05)		

Note. Values enclosed in parentheses represent mean square errors. Due to violation of sphericity assumption using Mauchly's W test ($p < .01$), the *df*'s in the within-subjects section were adjusted using Huynh-Feldt adjustments.
*$p < .05$. **$p < .01$.

Table 4.8: Post-hoc ANOVA results for proportions of IP breaks in Guoyu.

levels ($p < .05$ for DSP0, and $p < .01$ for DSP1 and DSP2). Hierarchically across three levels, post-hoc analyses using the Tukey's-*b* test showed that the proportion of IP breaks at P_0 reflects boundary strengths. There are proportionately more IP breaks at the DSP2 level than those at the DSP1 level, and there are in turn more IP breaks at the DSP1 level than those at the DSP0 level ($p < .05$).

In general, statistical test results match the observations above. There is a robust position effect in that IP breaks are more likely to occur at the boundary position. Hierarchically, the proportion of IP breaks at P_0 is indicative of discourse hierarchy.

4.7 Putonghua Results

As in Guoyu, there are also six possible values for generic breaks to assume in Putonghua—BI0, BI1, BI2, BI3, BI4, and BI5. Figure 4.3 shows the distribution of prosodic breaks in Putonghua. Like Guoyu, BI1 is the most common value for prosodic breaks. However, unlike Guoyu, there is also a considerable amount of BI2. In addition, there is a complementary distribution between P_0 and non-boundary syllable positions. BI0 and BI1 almost always occur at non-boundary positions while BI3, BI4, and BI5 almost always at P_0. Like Guoyu, the proportions of BI4 and BI5 increase as discourse disjuncture becomes higher.

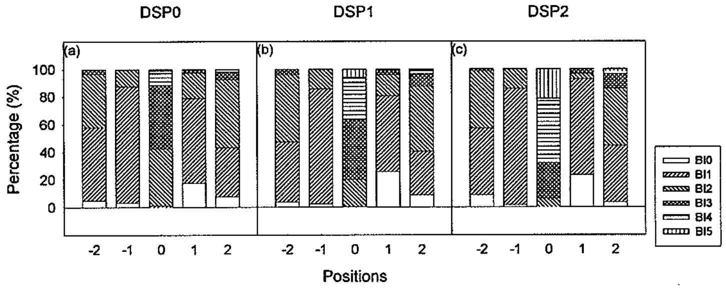

Figure 4.3: Bar graphs showing the percentages of BIs with regards to positions and discourse hierarchy in Putonghua. The layout of the graph is similar to that in Figure 4.1. (a) DSP0; (b) DSP1; (c) DSP2.

A post-hoc two-way Position (5) × Hierarchy (3) mixed design ANOVA on percentages of IP breaks (i.e., BI4 and BI5) was performed to confirm the above observations. Table 4.9 shows the results. Both of the main effects and the interaction effect are significant. Horizontally across different positions, post-hoc analyses using Bonferroni's adjustment showed that the proportion of IP breaks is higher at P_0 for the DSP1 and DSP2 levels ($p < .05$). Hierarchically across three levels, post-hoc analyses using the Tukey's-*b* test showed that the proportion of IP breaks at P_0 reflects boundary strengths. There are proportionately more IP breaks at the DSP2 level than those at the DSP1 and DSP0 levels ($p < .05$).

Source	*df*	*F*	η^2	*p*
		Between subjects		
Hierarchy (H)	2	18.46**	.64	.00
Error	21	(0.01)		
		Within subjects		
Position (P)	1.19	50.11**	.71	.00
P × H	2.39	14.25**	.58	.00
Error	25.05	(0.06)		

Note. Values enclosed in parentheses represent mean square errors. Due to violation of sphericity assumption using Mauchly's W test ($p < .01$), the *df*'s in the within-subjects section were adjusted using Huynh-Feldt adjustments.
**p* < .05. ***p* < .01.

Table 4.9: Post-hoc ANOVA results for proportions of IP breaks in Putonghua.

In general, statistical test results match the observations above. IP breaks are more likely to occur at the boundary position for DSP1 and DSP2 level boundaries. Hierarchically, there are in proportion more IP breaks at the boundary syllable when discourse disjuncture is at the DSP2 level as compared to disjunctures at lower levels.

4.8 Japanese Results

There are six possible values for breaks to assume in Japanese— BI0, BI1, BI2, BI3, BI4, and no BI. However, there are no BI0s in the corpus. Figure 4.4 shows the distribution of prosodic breaks in Japanese. Most of the syllables are without a BI. A complementary distribution exists between BI1, and BI3 and BI4. The latter group

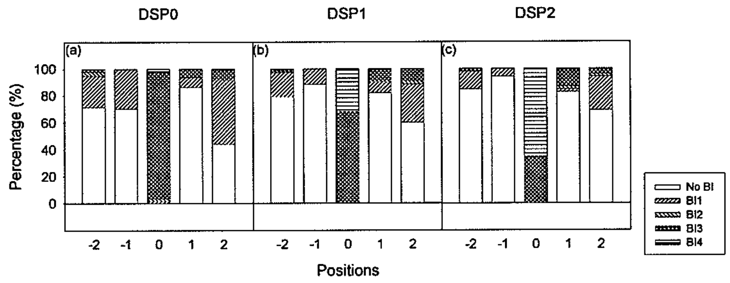

Figure 4.4: Bar graphs showing the percentages of BIs with regards to positions and discourse hierarchy in Japanese. The layout of the graph is similar to Figure 4.1. (a) DSP0; (b) DSP1; (c) DSP2.

appears mainly on P_0 while the former group appears mainly on non-boundary positions. There are not many BI2s at the boundary region. The proportion of BI4 at the boundary syllable increases as discourse hierarchy becomes higher.

A post-hoc two-way Position (5) × Hierarchy (3) mixed design ANOVA on percentages of IP breaks (i.e., BI3 and BI4) was performed to confirm the above observation. As shown in Table 4.10, only the position main effect is significant. The hierarchy main effect is near significant (p = .053) while the interaction effect is not significant. Horizontally across different positions, post-hoc analyses using Bonferroni's adjustment showed that the proportion of IP breaks is the highest at the boundary position ($p < .01$). Hierarchically across three levels, post-hoc analyses using the LSD test showed

Source	*df*	*F*	$\hat{\eta}^2$	*P*
		Between subjects		
Hierarchy (H)	2	3.39	.24	.05
Error	21	(0.01)		
		Within subjects		
Position (P)	2.31	692.95**	.97	.00
P × H	4.61	1.29	.11	.29
Error	48.45	(0.03)		

Note. Values enclosed in parentheses represent mean square errors. Due to violation of sphericity assumption using Mauchly's W test ($p < .01$), the *df*'s in the within-subjects section were adjusted using Huynh-Feldt adjustments.
*$p < .05$. **$p < .01$.

Table 4.10: Post-hoc ANOVA results for proportions of IP breaks in Japanese.

that proportions of IP breaks reflect hierarchy. There are in proportion more IP breaks at the DSP2 and DSP1 levels than at the DSP0 levels ($p < .05$).

In general, statistical test results match the observations above. There is a robust position effect. IP breaks are more likely to occur at the boundary position. There is also a slight hierarchy effect. Slightly more IP breaks are found at the DSP2 and DSP1 level than at the DSP0 level at the boundary position.

4.9 Cross-linguistic Comparisons

Figure 4.5 is a comparison of the percentages of IP breaks at the boundary position across four languages. As shown in the figure, languages have different proportions of IP breaks at this position, with English and Japanese being the highest and Guoyu and Putonghua being the lowest. However, all four languages show a consistent patterning reflecting hierarchy.

A post-hoc two-way Hierarchy (3) × Language (4) between-subject design ANOVA on IP break proportions was performed to confirm the above observations. Table 4.11 shows the results. Both of the main effects and the interaction effect are significant. Post-hoc Tukey's-*b* tests showed that at DSP0 and DSP1 levels, Japanese has the highest proportion of IP breaks at the boundary position, followed closely by English, and the two Mandarin varieties have the lowest proportions of IP breaks ($p < .05$). At the DSP2 level, however, there is no difference between English and Japanese, although the two languages still have higher proportions of IP breaks than Guoyu and Putonghua ($p < .05$).

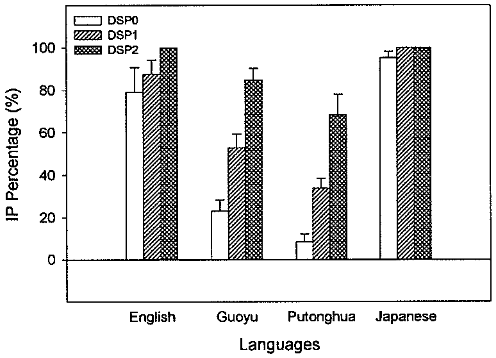

Figure 4.5: A bar graph illustrating the proportions of IP breaks at structural boundaries of all four languages. The *x*-axis represents languages and the *y*-axis represents percentage of IP breaks. The error bars indicate standard error.

Source	*df*	*F*	$\hat{\eta}^2$	*p*
		Between subjects		
Hierarchy (H)	2	39.28**	.48	.00
Language (L)	3	84.92**	.75	.00
H × L	6	2.97*	.18	.01
Error	84	(0.07)		

Note. Values enclosed in parentheses represent mean square errors.
$*p < .05$. $**p < .01$.

Table 4.11: Post-hoc ANOVA results for IP break proportions at the boundary position for all four languages.

Comparing Figures 4.1 through 4.4, one finds that the patterning is due to the fact that in both varieties of Mandarin, BI2 and BI3 (i.e., minor and major phrase boundaries, respectively) are allowed at the boundary syllable position, while for English and Japanese, this is not the case. Boundary position is almost always exclusively designated an IP boundary. On the other hand, BI2 is a dominant break category for Guoyu and Putonghua at the DSP0 level, and BI3 is a dominant category for the DSP0 and DSP1 levels in Putonghua.

4.10 Discussion

As shown in Sections 4.5–4.8, the patterning of prosodic breaks does not vary much cross-linguistically. In general, the position effect is fairly universal and accounts for a large amount of variance ($\hat{\eta}^2$ ranging from 71% in Putonghua to 97% in Japanese) For all four languages, proportions of IP breaks are higher at boundary positions than at non-boundary positions.

The hierarchy effect is more varied and accounts for a lesser amount of variance ($\hat{\eta}^2$ ranging from 24% in Japanese to 72% in Guoyu). There is a tendency for proportions of IP breaks at the boundary position to increase as discourse disjuncture becomes higher. In English, the level of discourse disjuncture is also reflected at P_1. Proportionately more IP breaks occur at this position when discourse disjuncture is high.

Cross-linguistically, one also finds that at P_0, English and Japanese have on average higher proportions of IP breaks than the two Mandarins. Guoyu in turn also has higher proportions of IP breaks than Putonghua at this position. This is due to the observation that Mandarin allows lower-level prosodic breaks (i.e., BI2 and BI3) at this position as compared to English and Japanese.

The qualitative nature of the prosodic break system (and thus the prosodic hierarchy) does not seem to reflect much of the language effect. One suspects that a quantitative measure such as those in the acoustic-phonetic domain might be a better candidate. Therefore, in the following chapters, two such measures, pitch range trend and duration are examined.

CHAPTER 5

RESULTS: PITCH PEAK TREND

Declination and pitch reset have been found to be a prevalent phenomenon within and across a prosodic domain, respectively. Since syntactic and prosodic boundaries often coincide with each other, it is not unlikely that declination and pitch reset might play a role in indicating the organization of discourse structures (Grosz & Hirschberg, 1992; Hirschberg & Nakatani, 1996; Ladd, 1988; Lehiste, 1975a; Swerts, 1997; Swerts & Geluykens, 1994). In this chapter, the relationship between topline declination and pitch reset and discourse hierarchy is examined. The reason why only topline declination is examined is because previous studies have shown that declination is more prominent with high than low tonal targets (Ladd, 1988).

5.1 Summary Statistics

Mean and median F_0 values of the four chosen points, P_{-2}, P_{-1}, P_1, and P_2 were calculated. As shown in Figure 5.1, all of the data points are near the $x = y$ line, indicating that the distribution of F_0 values is very close to normal. In general, male and female

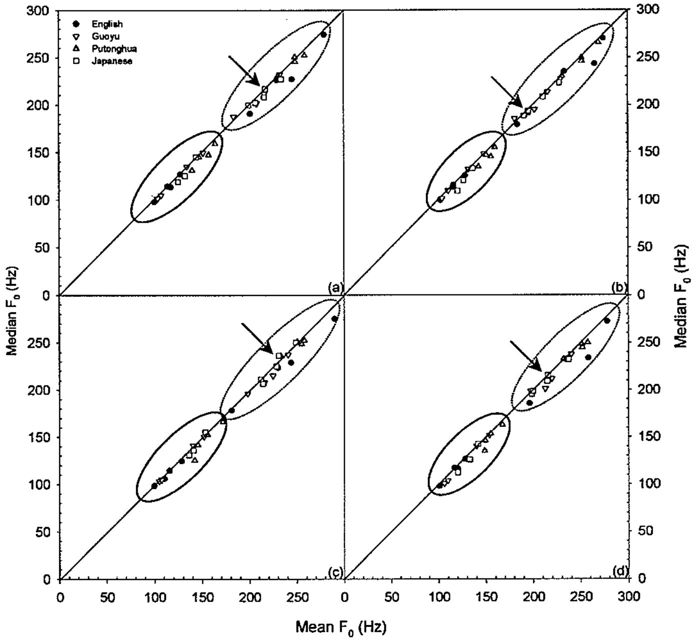

Figure 5.1: Scatter plots of the mean and median F_0 values of (a) P_{-2}, (b) P_{-1}, (c) P_1, and (d) P_2 of the 32 speakers of four languages. Each data point represents one speaker. The *x*-axes represent mean F_0 and the *y*-axes represent median F_0. Data points enclosed in a solid oval represent male speakers while those enclosed in a dashed oval represent female speakers. The arrows point to the data points of subject MS, who is a male, but his data points cluster with those of female rather than with other male speakers.

speakers are fairly segregated in terms of mean and median F_0. However, there are exceptions. Subject MS, a Japanese male speaker (indicated by the arrows), has means and medians closer to female than male speakers. There are also some language-specific trends that are worth noticing. English male speakers tend to cluster at the lower end of the distribution while English female speakers are more spread apart. Guoyu and Japanese speakers are the next lowest, regardless of gender. Putonghua speakers tend to display the highest pitch of all four languages.

5.2 Data Selection Criteria

Since analyses of pitch peak trends require at least two points across a structural boundary, there are cases where analyses are not possible due to lack of high tones or creaky voice quality. As a consequence, those cases were excluded from further analyses.

	English		Guoyu		Putonghua		Japanese		
	Male	Female	Male	Female	Male	Female	Male	Female	Total
DSP0	75	44	115	113	128	103	46	47	671
DSP1	92	70	108	104	117	81	57	51	680
DSP2	29	29	30	22	27	23	24	23	207
Total	196	143	253	239	272	207	127	121	1558

Table 5.1: Distribution of the number of cases at different discourse levels regarding pitch peak trend in all four languages.

Table 5.1 shows the distribution with regards to pitch trend analyses. As with the break index analyses, Japanese has the fewest cases of all four languages. For the number of cases each subject contributed, please see Appendix E.

5.3 Predictions

Since declination is likely to occur within a structural unit, and pitch reset across a unit, a reset is expected at the boundary region (Cooper & Sorensen, 1977; Fisher & Tokura, 1996; Grosz & Hirschberg, 1992; Hirschberg & Nakatani, 1996; Ladd, 1988; Lehiste, 1975a; Swerts, 1997; Swerts & Geluykens, 1994). It implies a jump in pitch peak values across a boundary as compared to that within. This effect will be termed the position effect, since the F_0 values are determined by their positions in time relative to the boundary. Also, it is possible that the magnitude of pitch reset is reflective of discourse hierarchy in that a bigger pitch reset (by a bigger pitch reset or a bigger declination or both) occurs at bigger discourse boundaries. In other words, a hierarchy effect is also expected. In addition, since physiological and sociolinguistic factors will also influence pitch height, a gender effect is examined as well. Finally, a language effect is also included in the analyses to entertain the hypothesis of language specificity.

5.4 Overall ANOVA

An overall Position (4) × Hierarchy (3) × Gender (2) × Language (4) four-way mixed design ANOVA was performed to test the above predictions, the within-subject variable being position, and the between-subject variables being language, gender, and hierarchy. There are four levels of language—English, Guoyu, Putonghua, and Japanese, three levels of hierarchy—DSP0, DSP1, and DSP2, two levels of gender--male and female, and four levels of position, which are the four high tones chosen from around the boundary region, two before the boundary (P_{-2} and P_{-1}), and two after (P_1 and P_2). Peak F_0 value is the dependent measure.

The ANOVA results are shown in Table 5.2. Three of the four main effects are significant. Hierarchy is the only factor that does not show a main effect. Three of the two-way effects and three of the three-way effects involving all four factors are also significant. The four-way interaction is significant as well.

The two-way, three-way, and four-way interactions imply that all four factors are at work in a complex manner. In the following sections, 5.5 through 5.8, detailed descriptions of the post-hoc test results are given and the pitch peak trend of each language is described.

5.5 English Results

Figure 5.2 shows the pitch peak trends with regards to disjuncture hierarchy in English. The *x*-axis represents positions relative to the boundary and the *y*-axis represents F_0 peak values in hertz. From the figure, one sees that pitch values are higher in female than male speakers, which was expected. Horizontally across different positions, one finds no consistent pattern of declination or pitch reset. If there is a pitch reset at the boundary position, and declination within a structural unit, one would expect the pitch value of P_1 to be higher than that of P_{-1}, and the pitch value of P_{-1} to be lower than that of

Source	*df*	*F*	$\hat{\eta}^2$	*p*
		Between subjects		
Hierarchy (H)	2	0.02	.00	.98
Gender (G)	1	3304.79**	.68	.00
Language (L)	3	113.25**	.18	.00
H × G	2	2.42	.00	.09
G × L	3	67.70**	.12	.00
L × H	6	0.90	.00	.49
H × G × L	6	2.17*	.01	.04
Error	1534	(2802.33)		
		Within subjects		
Position (P)	2.43	41.32**	.03	.00
P × H	4.86	28.61**	.04	.00
P × G	2.43	0.61	.00	.58
P × L	7.29	13.02**	.03	.00
P × H × G	4.86	1.26	.00	.28
P × G × L	7.29	2.78**	.01	.01
P × H × L	14.57	1.90*	.01	.02
P × G × H × L	14.57	1.72*	.01	.04
Error	3725.32	(657.29)		

Note. Values enclosed in parentheses represent mean square errors. Due to violation of sphericity assumption using Mauchly's W test ($p < .01$), the *df*'s in the within-subjects section were adjusted using Huynh-Feldt adjustments.

*$p < .05$. **$p < .01$.

Table 5.2: Overall ANOVA results for pitch peak trends.

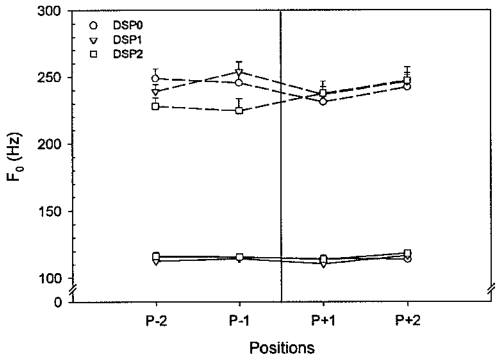

Figure 5.2: A line graph illustrating pitch peak trends at structural boundaries in English. The *x*-axis indicates position and the *y*-axis indicates F$_0$ values in hertz. The dashed vertical line indicates where the boundary is. P$_{-1}$ and P$_{-2}$ represent the two high tones closest to and before the boundary and P$_1$ and P$_2$ represent the two high tones closest to and after the boundary. The solid lines represent male speakers, and the dashed lines represent female speakers. The error bars represent standard error. Only the upper part is shown to avoid clutter.

P$_{-2}$. This pattern is not observed. Instead, one finds the pitch trend in male speakers is almost flat. If anything, it *rises* within boundary and *declines* across instead. The pitch peak trend in female speakers is not uniform. The F$_0$ contour of the pre-boundary unit either declines or rises, and that of the post-boundary unit tends to start low and rises.

A post-hoc Position (4) × Gender (2) × Hierarchy (3) three-way mixed design ANOVA was performed to confirm the above observations. As shown in Table 5.3, main effects of position and gender are significant. The interaction between position and hierarchy is also significant. Three way interaction is near-significant (p = .051). A post-hoc *t*-test showed that female speakers have significantly higher pitch values than female speakers at all four positions (p < .01). Horizontally across positions, three planned paired *t*-test comparisons were performed—between P$_{-2}$ and P$_{-1}$, between P$_{-1}$ and

Source	*df*	*F*	$\hat{\eta}^2$	*P*
		Between subjects		
Hierarchy (H)	2	0.48	.00	.62
Gender (G)	1	1698.56**	.84	.00
H × G	2	1.21	.01	.30
Error	333	(2556.63)		
		Within subjects		
Position (P)	2.87	3.11*	.01	.03
P × H	5.74	2.88**	.02	.01
P × G	2.87	0.86	.00	.46
P × H × G	5.74	2.12	.01	.05
Error	954.82	(756.09)		

Note. Values enclosed in parentheses represent mean square errors. Due to violation of sphericity assumption using Mauchly's W test (p < .01), the *df*'s in the within-subjects section were adjusted using Huynh-Feldt adjustments.

**p* < .05. ***p* < .01.

Table 5.3: Post-hoc ANOVA results for pitch peak trends in English.

P_l, and between P_l and P_2. Results showed that for female speakers at the DSP0 level, and male and female speakers at the DSP1 level, P_{-1} has higher pitch values than P_l (male: $p < .01$; female: $p < .05$). In addition, male speakers also showed a P_2 that is higher in pitch than P_l at the DSP1 and DSP2 levels (DSP1: $p < .01$; DSP2: $p = .07$), while female speakers showed higher pitch values at P_{-1} than P_{-2} at the DSP1 level ($p < .05$). Hierarchically, post-hoc Tukey's-*b* tests showed that P_{-1} has a higher pitch value at the DSP1 level than that at the DSP2 level ($p < .05$) in female speech. All the other trends are nonsignificant.

Statistical analyses confirmed the observations above. Male speakers in general have lower pitch values than female speakers. There is a minor position effect. Pitch contours tend to end high in the pre-boundary unit and start low in a new structural unit. There is no pitch reset and no consistent hierarchy effect was found.

5.6 Guoyu Results

Figure 5.3 shows the pitch peak trends with regards to disjuncture hierarchy in Guoyu. The layout of the figure is similar to that of Figure 5.2. As in English, Guoyu male speakers tend to have lower pitch values than female speakers. However, different from English, one finds a more consistent patterning across different positions. Except for the DSP0 level, there is declination within the preboundary unit and reset across the boundary. This is true with both male and female speakers. At the DSP0 level, the pitch trend is almost monotonously flat, especially for female speakers. Male speakers tend to have a slight rise within boundaries and a slight fall across boundaries at this level. Hierarchically, one sees that female speakers have a higher degree of declination within the preboundary unit when the boundary disjuncture is high. In other words, pitch trends decline to a lower level with bigger boundaries. The amount of reset is also higher. In male speakers, only the magnitude of pitch reset is reflective of hierarchy.

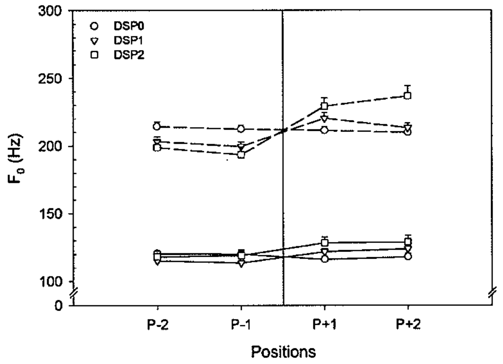

Figure 5.3: A line graph illustrating the pitch peak trends at structural boundaries in Guoyu. The layout is similar to that in Figure 5.2. The dashed vertical line indicates where the boundary is. The solid lines represent male speakers, and the dashed lines represent female speakers. The error bars represent standard error. Only the upper part is shown to avoid clutter.

A post-hoc Position (4) × Gender (2) × Hierarchy (3) three-way mixed design ANOVA was performed to confirm the above observations. Table 5.4 shows the results. Main effects of position and gender are significant. In addition, two of the two-way interaction and the three-way interaction effects are significant. Post-hoc *t*-tests concerning the gender effect showed that female speakers have higher pitch values than male speakers at all four positions ($p < .01$). Horizontally across positions, three planned paired *t*-tests were performed—between P_{-2} and P_{-1}, between P_{-1} and P_1, and between P_1 and P_2. At the DSP0 level, male speakers showed that P_{-1} has higher pitch values than P_1 ($p < .01$), while for female speakers, there is no such difference. At the DSP1 and DSP2 levels, P_1 always have higher pitch values than P_{-1}, regardless of gender (male at

Source	*df*	*F*	$\hat{\eta}^2$	*p*
		Between subjects		
Hierarchy (H)	2	0.93	.00	.40
Gender (G)	1	1078.83**	.69	.00
H × G	2	0.19	.00	.83
Error	486	(2428.67)		
		Within subjects		
Position (P)	2.22	38.50**	.07	.00
P × H	4.43	21.91**	.08	.00
P × G	2.22	11.14**	.02	.00
P × H × G	4.43	4.65**	.02	.00
Error	1076.63	(430.49)		

Note. Values enclosed in parentheses represent mean square errors. Due to violation of sphericity assumption using Mauchly's W test ($p < .01$), the *df*'s in the within-subjects section were adjusted using Huynh-Feldt adjustments.

*$p < .05$. **$p < .01$.

Table 5.4: Post-hoc ANOVA results for pitch peak trends in Guoyu.

DSP2: $p < .05$; $p < .01$ for the others). In addition, one also finds that at the DSP1 level, P_1 has higher pitch values than P_2 for the female speakers ($p < .05$).

Hierarchically speaking, post-hoc Tukey's-*b* tests indicated that male speakers showed an effect at the P_1 position. Pitch values at the DSP2 levels are higher than those at the DSP0 level ($p < .05$). In female speech, pitch values at the DSP2 level are higher than those at the other levels at the P_1 and P_2 positions ($p < .05$). At the P_{-1}, pitch values are higher at the DSP0 level than the DSP2 level ($p < .05$). Using the post-hoc LSD test, one also finds that P_{-2} pitch values are higher at the DSP0 level than the DSP1 level ($p < .05$).

In general, statistical results confirmed the observations above. Male speakers did not have as high pitch values as female speakers. However, both genders show a significant reset effect across a structural boundary, especially at higher discourse disjunctures. The degree of pitch reset corresponds positively to discourse hierarchy. Bigger boundaries show a bigger pitch reset. Female speakers also use the degree of declination as a hierarchy indicator.

5.7 Putonghua Results

Figure 5.4 shows the pitch peak trends with regards to disjuncture hierarchy in Putonghua. The layout of the figure is similar to that of Figures 5.2 and 5.3. As in English and Guoyu, male speakers in Putonghua have lower pitch values than female speakers. However, unlike the two languages, pitch trends are more drastic in male than female speech. Horizontally across different positions, one again sees a pitch reset at the DSP1 and DSP2 levels, but not at the DSP0 level. This is true of both genders. At the DSP0

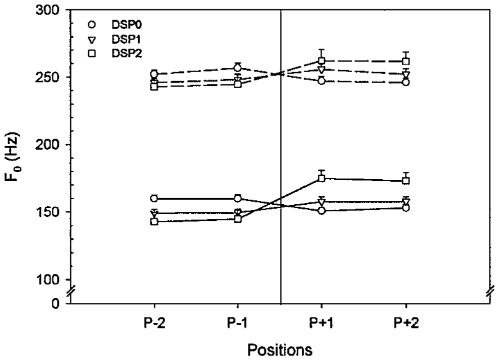

Figure 5.4: A line graph illustrating pitch peak trends at structural boundaries in Putonghua. The layout is similar to that in Figure 5.2. The dashed vertical line indicates where the boundary is. The solid lines represent male speakers, and the dashed lines represent female speakers. The error bars represent standard error. Only the upper part is shown to avoid clutter.

level, there is a downward pitch trend between P_{-1} and P_1. Hierarchically, both genders showed that the magnitude of declination and pitch reset is positively correlated with hierarchy. Bigger disjunctures have bigger reset and declination.

A post-hoc Position (4) × Gender (2) × Hierarchy (3) three-way mixed design ANOVA was performed to confirm the above observations. Table 5.5 shows the results. Two of the main effects, position and gender, and a two-way interaction effect, Position × Hierarchy, are significant. Post-hoc *t*-tests showed that male speakers have significantly

lower pitch values than female speakers at all four positions ($p < .01$). Horizontally across positions, three planned paired *t*-tests were performed—between P_{-2} and P_{-1}, between P_{-1} and P_1, and between P_1 and P_2. Since the interaction effect involving gender is not significant, data from male and female speakers were combined for the post-hoc analyses regarding position and hierarchy. For the position effect, at the DSP0 level, P_{-1} has higher pitch values than P_1 ($p < .01$). At the DSP1 and DSP2 levels, speakers showed a P_1 that is higher in pitch than P_{-1} ($p < .01$).

Source	*df*	*F*	η^2	*P*
		Between subjects		
Hierarchy (H)	2	0.49	.00	.62
Gender (G)	1	1161.15**	.71	.00
H × G	2	0.15	.00	.86
Error	473	(2357.37)		
		Within subjects		
Position (P)	1.97	10.86**	.02	.00
P × H	3.93	14.45**	.06	.00
P × G	1.97	0.98	.00	.37
P × H × G	3.93	0.49	.00	.74
Error	929.98	(932.39)		

Note. Values enclosed in parentheses represent mean square errors. Due to violation of sphericity assumption using Mauchly's W test ($p < .01$), the *df*'s in the within-subjects section were adjusted using Huynh-Feldt adjustments.

*$p < .05$. **$p < .01$.

Table 5.5: Post-hoc ANOVA results for pitch peak trends in Putonghua.

Hierarchically, post-hoc Tukey's-*b* tests showed that P_1 pitch values are higher at the DSP2 level as compared to those at the DSP0 levels ($p < .05$), while P_2 pitch values are the highest at the DSP2 level ($p < .05$). Post-hoc LSD tests also showed that P_{-1} and P_{-2} pitch values are higher at the DSP0 level than those at the DSP1 level ($p < .05$).

In general, statistical analyses confirmed the observations above. Male speakers have lower pitch values than female speakers. However, they also have clearer pitch reset and declination cues. Pitch reset is found at the DSP1 and DSP2 levels, although only that for the male speakers is statistically significant. At the DSP0 level, there is no pitch reset. Hierarchically, the magnitude of the pitch reset and declination is also reflective of discourse disjuncture. Both genders showed the trend, but there is a bigger effect in male speech.

5.8 Japanese Results

Figure 5.5 shows the pitch peak trends with regards to disjuncture hierarchy in Japanese. The layout of the figure is similar to that of Figures 5.2, 5.3, and 5.4. As in the other languages, male speakers have lower pitch values than female speakers. However, male speakers have more drastic pitch contours than female speakers, as in Putonghua. Horizontally across different positions, there is a consistent pattern of declination within a structural unit and pitch reset across a boundary in both genders. Hierarchically, the degree of pitch reset in male speech and declination in female speech is reflective of discourse hierarchy.

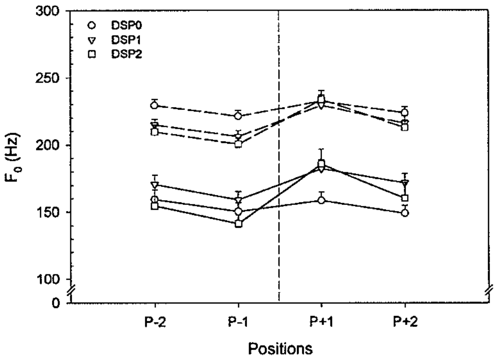

Figure 5.5: A line graph illustrating pitch peak trends at structural boundaries in Japanese. The layout is similar to that in Figure 5.2. The dashed vertical line indicates where the boundary is. The solid lines represent male speakers, and the dashed lines represent female speakers. The error bars represent standard error. Only the upper part is shown to avoid clutter.

A post-hoc Position (4) × Gender (2) × Hierarchy (3) three-way mixed design ANOVA was performed to confirm the above observations. As shown in Table 5.6, the main position and gender effects are significant. Two of the two-way interactions involving hierarchy are also significant. Post-hoc *t*-tests showed that female speakers again have higher pitch values than male speakers ($p < .01$). Horizontally across different positions, three planned paired *t*-tests were performed—between P_{-2} and P_{-1}, between P_{-1} and P_1, and between P_1 and P_2. Since there is no Position × Gender × Hierarchy effect, data from both genders were combined for the post-hoc analyses regarding position. Results showed that pitch values at P_{-2} are always higher than those at P_{-1}, those at P_1 are always higher than those at P_2, and those at P_1 are always higher than those at P_{-1}, regardless of the level of hierarchy ($p < .01$). Hierarchically, post-hoc Tukey's-*b* tests showed that for male speakers, P_1 pitch values are higher at the DSP2 level than those at the DSP0 level ($p < .05$), while for female speakers, P_{-2} and P_{-1} pitch values are higher at the DSP0 level than at the DSP2 levels ($p < .05$).

Source	*df*	*F*	$\hat{\eta}^2$	*p*
		Between subjects		
Hierarchy (H)	2	0.63	.01	.54
Gender (G)	1	149.37**	.38	.00
H × G	2	3.78*	.03	.02
Error	242	(4760.53)		
		Within subjects		
Position (P)	2.50	38.07**	.14	.00
P × H	4.99	5.43**	.04	.00
P × G	2.50	0.16	.00	.89
P × H × G	4.99	0.36	.00	.88
Error	604.19	(655.55)		

Note. Values enclosed in parentheses represent mean square errors. Due to violation of sphericity assumption using Mauchly's W test ($p < .01$), the *df*'s in the within-subjects section were adjusted using Huynh-Feldt adjustments.

*$p < .05$. **$p < .01$.

Table 5.6: Post-hoc ANOVA results for pitch peak trends in Japanese.

In general, statistical analyses confirmed the above observations. Female speakers have higher pitch values than male speakers. Regardless of gender, declination and pitch reset are prominent within and across a structural boundary, respectively. Moreover, male speakers use the degree of pitch reset as a way to reflect discourse hierarchy while female speakers use the degree of declination to do so.

5.9 Cross-linguistic Comparisons

As shown in the previous sections, 5.5 to 5.8, languages have different pitch peak trends with regards to discourse boundary hierarchy. In English, both the boundary and hierarchy effects are not reflected by pitch reset and declination. However, in Guoyu, Putonghua, and Japanese, there is some evidence showing that pitch peak patterning serves as an indicator for boundary location and disjuncture strength.

Figure 5.6 is a comparison of the pitch values of P_{-1} and P_1 in the four languages. One sees that although in all four languages, male speakers are consistently lower in pitch than female speakers, the magnitude of the gender difference is language-dependent. The average pitch peak values between male and female speakers are the most different in English, and most alike in Japanese. Also, when comparing within gender, one finds that English and Putonghua have higher pitch values than Guoyu and Japanese for female speakers while Putonghua and Japanese have higher pitch values than English and Guoyu for male speakers. In terms of pitch reset, Guoyu female speakers show a more prominent effect than male speakers, while in Putonghua and Japanese, it is the opposite. Male speakers show a bigger effect than female speakers.

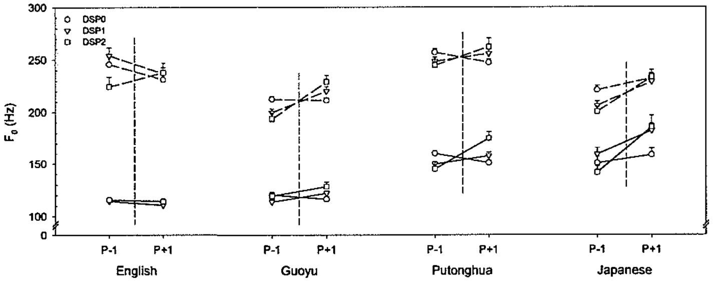

Figure 5.6: A line graph illustrating pitch peaks of P_{-1} and P_1 at structural boundaries in the four languages. The *x*-axis represents languages and the *y*-axis represents F_0 values in hertz. The solid lines represent male speakers, and the dashed lines represent female speakers. The error bars represent standard error. Only the upper part is shown to avoid clutter. The dashed vertical lines indicate where the boundaries are.

A post-hoc Position (2) × Gender (2) × Language (4) three-way mixed design ANOVA was performed to confirm the above observations. As shown in Table 5.7, all of the main effects are significant. Two of the two-way interactions involving language are also significant. In addition, the three-way interaction is significant. Post-hoc analyses using Tukey's-*b* tests showed that for male speakers, Putonghua and Japanese have higher P_{-1} pitch values than Guoyu and English ($p < .05$). For the P_1 position, Japanese has the highest pitch value, followed by Putonghua, Guoyu, and English ($p < .05$). With female speakers, Putonghua and English have the higher P_{-1} pitch values than Guoyu

Source	*df*	*F*	$\hat{\eta}^2$	*p*
		Between subjects		
Language (L)	3	162.47**	.24	.00
Gender (G)	1	3844.82**	.71	.00
L × G	3	69.98**	.12	.00
Error	1550	(1617.07)		
		Within subjects		
Position (P)	1	36.21**	.02	.00
P × L	3	29.26**	.05	.00
P × G	1	0.07	.00	.79
P × L × G	3	3.91**	.01	.01
Error	1550	(652.71)		

Note. Values enclosed in parentheses represent mean square errors.

*$p < .05$. **$p < .01$.

Table 5.7: Post-hoc ANOVA results for pitch peak values at P_{-1} and P_1 for all four languages.

and Japanese. At the P_1 position, Putonghua has the highest pitch value, followed by English and Japanese. Guoyu has the lowest pitch value of all four languages ($p < .05$).

The statistical analyses confirmed the observations above. For female speakers, English and Putonghua have the highest pitch values at the boundary region while for male speakers, Putonghua and Japanese have the highest pitch values.

5.10 Discussion

Unlike prosodic breaks, there is more cross-linguistic variance in pitch peak trends with regards to boundary cues in that not all languages take advantage of this parameter in a systematic manner. In English, for example, there is no consistent pattern of pitch reset and declination indicating boundary location and boundary strength that is of observational or statistical significance. On the other hand, in Guoyu, Putonghua, and Japanese, there is a general tendency for a pitch reset to occur across a structural boundary, although in Guoyu and Putonghua, the tendency only appears at the DSP1 and DSP2 levels.

The degree of reset and declination is also reflective of boundary strength. Bigger boundaries are more likely to have bigger reset and declination effects. This is true of female speakers in Guoyu and both genders in Putonghua. Male speakers in Guoyu only uses pitch reset to reflect hierarchy. In Japanese however, there is a gender difference in this regard. Female speakers tend to use the degree of declination to indicate hierarchy, while male speakers tend to use the degree of reset to do so.

In addition to discourse boundary size, the degree of pitch reset and its relation with discourse boundaries are also modified by gender. English and Guoyu showed more obvious use of pitch peak trends in female speech. This results in the fact that female speakers tend to show a more significant boundary effect, especially in Guoyu. For Putonghua and Japanese, however, the tendency is reversed. Male speakers showed a more prominent pitch reset effect than female speakers.

Finally, one also sees that gender difference is modified by language. In terms of pitch values, English speakers showed the most gender difference while Japanese speakers showed the least. This is due to the fact that English female speakers and Japanese male speakers have relatively high pitch values while English male speakers and Japanese female speakers have relatively low pitch values at the boundary region.

CHAPTER 6

RESULTS: SYLLABLE DURATION AND SYLLABLE ONSET INTERVAL

Syllable duration is influenced by many factors, including segmental effects such as phone types, suprasegmental effects such as stress, tone, and pitch accent, structural effects such as final lengthening at word, phrasal, clausal, and discourse boundaries, and pragmatic effects such as genre, rate, and style (Campbell & Isard, 1991; Cooper, Paccia, & Lapointe, 1978; Fowler, Levy, & Brown 1998; Huggins, 1974; Katz, Beach, Jenouri, & Verma, 1996; Klatt, 1975; Koopmans-van Beinum & Donzel, 1996; Kreiman, 1982; Lehiste, 1971, 1973, 1975b, 1979a, 1979b; Lehiste, Olive, & Streeter, 1976; Lehiste & Wang, 1977; Nakatani, O'Conner, & Aston, 1981; Oller, 1973; Quené, 1992, 1993; Scott, 1982; Shen, 1992; Streeter, 1978). On the other hand, silent pause duration and thus syllable onset intervals (SOIs) are mainly affected by structural and pragmatic parameters only (Cooper, Paccia, & Lapointe, 1978; Fon & Johnson, 2000; Hirschberg & Nakatani, 1996; Katz, Beach, Jenouri, & Verma, 1996; Ladd, 1988; Lehiste, 1979a, 1979b; Lehiste & Wang, 1977; Maclay & Osgood, 1959; O'Malley, Kloker, & Dara-Abrams, 1973; Scott, 1982; Swerts, 1998; Swerts & Geluykens, 1994). This chapter discusses how syllable duration and SOIs reflect discourse hierarchy in English, Mandarin and Japanese.

6.1 Summary Statistics

Each speaker has his own speaking style that can be easily recognized by people who are acquainted with the speaker. Similarly, speakers of the same language/dialect also show some homogeneity in style to a non-native ear. If speaking rate is one indicator for style, insofar as syllable duration and SOI reflect speaking rate, one would expect it to differ from speaker to speaker, and from language to language. Figure 6.1a is a scatter plot of the mean and median syllable duration of the 32 speakers of the four languages. Syllables here include only regular syllables (i.e., excluding filled pauses, such as, *um* in English, *uN* in Mandarin, and *anO* in Japanese). All data points are below the $x = y$ line, indicating that the distribution of syllable durations for each speaker is skewed to the right. In other words, Most syllables are short, with a few very long syllables, which is what one would expect from natural speech. As predicted above, the scatter plot shows a wide range of possible mean and median syllable durations, from the minimum mean 133.31 ms, median 116.04 ms to the maximum mean 226.02 ms, median 193.14 ms. In addition, each language also seems to have its own preferred mean and median syllable duration. Of the four, English has the longest syllable duration, followed by Guoyu and Putonghua, with Japanese having the shortest syllables.

Figure 6.1b is a scatter plot of the mean and median SOIs of the same 32 speakers. As in Figure 6.1a, only regular syllables were included. Again, all the data points are below the $x = y$ line, indicating a distribution of SOI durations for each speaker that is skewed to the right. Since an SOI is calculated by summing the duration of a syllable and any following silent pauses, average SOI should by definition be larger than average

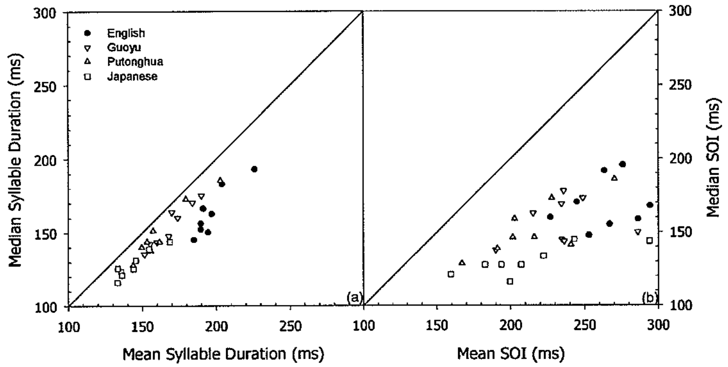

Figure 6.1: Scatter plots of the mean and median (a) syllable duration and (b) syllable onset intervals (SOIs) of the 32 speakers of the four languages. Each data point represents one speaker.

syllable duration. However, as shown in Figure 6.1, the increase is more obvious in the mean than the median, which stays more or less the same. This implies that the distribution is more spread out for SOIs, and longer syllables are more likely to be followed by longer pauses. Another point worth noticing is that the distribution of SOIs of the four languages is not as neatly segregated. Although Japanese still has the shortest intervals and English the longest, there is more overlap among the languages.

6.2 Data Selection Criteria

Figure 6.2 shows the distribution of different types of boundary syllables summed across all 32 subjects of the four languages/dialects. It is clear that boundary syllables are not evenly distributed across types. For English, Guoyu, and Putonghua, the most common boundary syllable types are either syllables by themselves or syllables that are followed by unfilled pauses. In Japanese, boundary syllables seldom occur alone. They are often followed by unfilled pauses or unfilled pauses plus filled pauses. Since preliminary studies seem to indicate that what follows a syllable has an effect on syllable

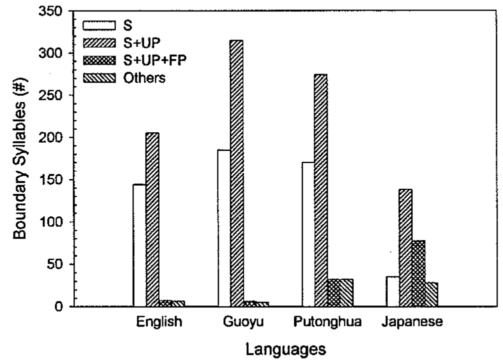

Figure 6.2: A bar graph showing the distribution of boundary syllable types. S: syllable only; S+UP: syllable followed by an unfilled pause; S+UP+FP: syllable followed by an unfilled plus a filled pause.

duration itself, in order to facilitate cross-linguistic comparisons, for all subsequent analyses, only cases containing boundary syllables that are followed by unfilled pauses are included. This is to maintain a more homogeneous set of data and facilitate cross-linguistic comparisons, despite the spontaneous nature of the study.

Table 6.1 shows the distribution of the number of cases each language has at different discourse levels using the selection criteria above. As one sees from the table, there are more cases of the DSP1 level than of the other two levels. Appendix F shows the number of cases each subject contributed.

6.3 Predictions

Duration has long been noted to be a strong indicator of sentential/discourse boundaries (Fougeron & Keating, 1997; Goldman-Eisler, 1972; Kreiman, 1982; Lehiste, 1975b, 1979a, 1979b; Lehiste & Wang, 1977; Oller, 1973; Shen, 1992; Swerts & Geluykens, 1994). Syllable duration and SOI tend to increase at boundary positions, a phenomenon that is generally known as final lengthening. In this study, the existence of such an increase will be referred to as a position effect since the closer a syllable or an SOI is to the sentential/discourse boundary, the more likely it is to be lengthened. The second effect examined is a hierarchy effect. One expects the degree of discourse disjuncture to be mirrored by the magnitude of acoustic cues. In other words, the bigger the boundary, the longer the syllable and pause duration at boundary positions. Finally, based on Figure 6.1, one might also expect a language effect in that different languages/dialects might show a different pattern in final lengthening.

	English	Guoyu	Putonghua	Japanese	Total
DSP0	47	103	115	35	300
DSP1	114	161	123	70	468
DSP2	44	51	36	33	164
Total	205	315	274	138	932

Table 6.1: Distribution of the number of cases each language has at different discourse levels.

6.4 Overall ANOVA

Based on the selection criteria above, the boundary syllables included can be followed either by another syllable or a silent pause. One finds that whether a syllable is followed by an unfilled pause is in part influenced by discourse levels. Table 6.2 shows

	English		Guoyu		Putonghua		Japanese	
(%)	S	S+UP	S	S+UP	S	S+UP	S	S+UP
DSP0	61.83	35.88	56.07	43.10	45.45	47.52	23.30	33.98
DSP1	31.21	65.90	21.76	74.54	26.67	58.57	7.50	58.33
DSP2	15.52	75.86	7.14	91.07	7.14	64.29	3.64	60.00

Table 6.2: Proportions of boundary syllables that are immediately followed by another syllable (S) and those that are followed by a silent pause (S+UP).

the proportions of boundary syllables that are immediately followed by another syllable and those that are followed by a pause. As shown in the table, unfilled pauses are more likely to occur after a boundary syllable when discourse disjuncture is high. The actual proportions vary among languages.

For statistical analyses, two overall Position (5) × Hierarchy (3) × Language (4) three-way mixed design ANOVAs were performed to test the above predictions, the within-subject variable being position, and the between-subject variables being language and hierarchy. There are four levels of language—English, Guoyu, Putonghua, and Japanese, three levels of hierarchy—DSP0, DSP1, and DSP2, and five levels of position, which are the five syllables selected from around the boundary region, three before the boundary, and two after. Two dependent variables, syllable duration and SOI, were tested.

ANOVA results with syllable duration as the dependent variable are shown in Table 6.3. Two main effects, language and position, are significant. The three-way interaction and all the two-way interactions are also significant.

ANOVA results with SOI as the dependent variable are shown in Table 6.4. All three main effects are significant. Two of the two-way interactions are also significant and the three-way interaction is near significant criterion ($p = .10$).

The significant two-way interaction effects and the common trend in the three-way interactions on both of the dependent variables suggest that all three factors are at work in lengthening. In the following sections, 6.5 through 6.8, detailed descriptions of the post-hoc test results are given and the durational patterning of each language is described.

Source	*df*	*F*	$\hat{\eta}^2$	*p*
		Between subjects		
Language (L)	3	96.01**	.24	.00
Hierarchy (H)	2	0.23	.00	.80
L × H	6	5.75**	.04	.00
Error	920	(0.01)		
		Within subjects		
Position (P)	3.42	153.61**	.14	.00
P × L	10.25	34.44**	.10	.00
P × H	6.83	4.09**	.01	.00
P × H × L	20.49	2.56**	.02	.00
Error	3141.67	(0.04)		

Note. Values enclosed in parentheses represent mean square errors. Due to violation of sphericity assumption using Mauchly's W test ($p < .01$), the *df*'s in the within-subjects section were adjusted using Huynh-Feldt adjustments.

*$p < .05$. **$p < .01$.

Table 6.3: Overall ANOVA results for syllable duration.

Source	*df*	*F*	η^2	*p*
		Between subjects		
Language (L)	3	17.98**	.06	.00
Hierarchy (H)	2	27.91**	.06	.00
L × H	6	0.75	.01	.61
Error	920	(0.09)		
		Within subjects		
Position (P)	2.34	1051.73**	.53	.00
P × L	7.01	10.92**	.03	.00
P × H	4.67	17.76**	.04	.00
P × H × L	14.01	1.51	.01	.10
Error	2148.63	(0.14)		

Note. Values enclosed in parentheses represent mean square errors. Due to violation of sphericity assumption using Mauchly's W test ($p < .01$), the *df*'s in the within-subjects section were adjusted using Huynh-Feldt adjustments.

$^*p < .05$. $^{**}p < .01$.

Table 6.4: Overall ANOVA results for SOI duration.

6.5 English Results

Figure 6.3 shows the patterning of syllable and pause duration at various discourse boundaries in English. The *x*-axis indicates position. The bars in the upper part of the graph (corresponding to the left *y*-axis) indicate syllable duration while those in the

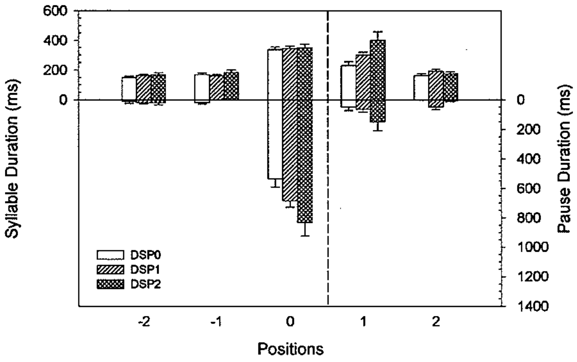

Figure 6.3: A bar graph illustrating the patterns of syllable and pause duration at structural boundaries in English. The *x*-axis shows syllable positions relative to the boundary. '0' indicates boundary syllables located right before the relevant boundaries, '-1' and '-2' indicate syllables that are one and two syllables before boundary syllables, respectively, and '1' and '2' indicate syllables that are one and two syllables after the relevant boundaries, respectively. In other words, the boundary is between '0' and '1', as indicated by the dashed line. The left *y*-axis indicates syllable duration and the right *y*-axis indicates pause duration. The error bars indicate standard error.

lower part (corresponding to the right *y*-axis) indicate pause duration. In other words, the total length of the bars (combining the upper and the lower parts) indicates SOI.

From the figure, one sees that although syllable duration is lengthened at Position 0 (hereafter P_0) as compared to that at Positions -1 and -2 (hereafter P_{-1} and P_{-2}, respectively), there is also lengthening at Position 1 (hereafter P_1), especially at Levels DSP1 and DSP2. In fact, at the DSP2 level, the highest degree of lengthening falls on P_1 instead of P_0. This initial lengthening effect is somewhat unexpected. On the other hand, SOI patterning follows what was predicted in that the highest degree of lengthening occurs at P_0. However, there is still a slight lengthening effect at P_1, especially at Level DSP2. Both initial and final lengthening effects seem to reflect boundary strength.

Two post-hoc Position (5) × Hierarchy (3) two-way mixed design ANOVAs were performed to confirm the above observations. Table 6.5 shows the results with syllable duration. Both of the main effects and the interaction effect are significant. Horizontally across different positions, post-hoc analyses using Bonferroni's adjustments showed that at the DSP0 level, syllables at P_0 are the longest ($p < .05$) while at DSP1 and DSP2 levels, syllables at both P_0 and P_1 are the longest ($p < .01$). The two are not significantly different from each other, however. On the other hand, post-hoc analyses using the Tukey's-*b* test concerning the hierarchy effect showed that syllables at P_0 do not reflect boundary strength but those at P_1 do. Syllables at the DSP2 level at this position are significantly longer than those at DSP1 and DSP0 levels ($p < .05$).

A similar post-hoc two-way ANOVA concerning SOIs was also performed. As shown in Table 6.6, both of the main effects are again significant. The interaction effect is also significant. Horizontally across the five positions, post-hoc tests using Bonferroni's

Source	*df*	*F*	η^2	*p*
		Between subjects		
Hierarchy (H)	2	5.41**	.05	.01
Error	202	(0.02)		
		Within subjects		
Position (P)	2.63	59.60**	.23	.00
P × H	5.26	2.80*	.03	.02
Error	531.27	(0.04)		

Note. Values enclosed in parentheses represent mean square errors. Due to violation of sphericity assumption using Mauchly's W test ($p < .01$), the *df*'s in the within-subjects section were adjusted using Huynh-Feldt adjustments.

**p* < .05. ***p* < .01.

Table 6.5: Post-hoc ANOVA results for syllable duration in English.

adjustments showed that at all three DSP levels, SOIs at P_0 are the longest ($p < .01$). In addition, at DSP1 and DSP2 levels, SOIs at P_1 are also significantly longer than those at P_{-2}, P_{-1}, and P_2, although they are still shorter than those at P_0 ($p < .05$). Hierarchically, Tukey's-*b* test showed that SOI duration at both P_0 and P_1 mirrors boundary strength. SOIs at the DSP2 level are significantly longer than those at other levels ($p < .05$).

The initial lengthening effect, which reflects boundary strength, was unexpected. One possible explanation is that the effect might be related to pitch accents. If higher-level discourse segments are more likely to have accented full referring NPs as subjects, and accented syllables are more likely to be lengthened, then lengthening in position P_1 might be a reflection of the greater likelihood of having a subject with initial

Source	*df*	*F*	η^2	*p*
		Between subjects		
Hierarchy (H)	2	6.83**	.06	.00
Error	202	(0.13)		
		Within subjects		
Position (P)	2.45	217.64**	.52	.00
P × H	4.89	2.96*	.03	.01
Error	494.15	(0.42)		

Note. Values enclosed in parentheses represent mean square errors. Due to violation of sphericity assumption using Mauchly's W test ($p < .01$), the *df*'s in the within-subjects section were adjusted using Huynh-Feldt adjustments.

*$p < .05$. **$p < .01$.

Table 6.6: Post-hoc ANOVA results for SOI duration in English.

Position	DSP0	DSP1	DSP2	Total
P_0				
accented	24	57	23	104
unaccented	23	57	21	101
P_1				
accented	16	63	28	107
unaccented	31	51	16	98

Table 6.7: Distribution of accented and unaccented syllables at P_0 and P_1 with regards to discourse hierarchy.

accent in that position. To test this explanation, syllables at P_0 and P_1 were subcategorized by the existence of pitch accents. Table 6.7 shows the distribution and chi-square tests showed that the distribution is significant [$\chi^2(2) = 8.95, p < .05$].

Figure 6.4 shows that pitch accents lengthen syllable and pause duration in general, regardless of positions. However, it is more effective in lengthening syllables than SOIs, and at P_1 than P_0. The initial strengthening effect observed above seems to be due solely to the existence of pitch accents. Unaccented syllables are lengthened only at P_0, DSP-final positions.

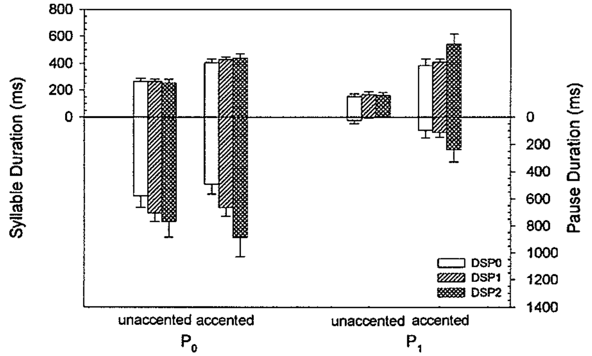

Figure 6.4: A bar graph illustrating the duration patterns of accented and unaccented syllables and SOIs at P_0 and P_1. The *x*-axis indicates positions, and the *y*-axes indicate syllable (left) and pause duration (right). The error bars indicate standard error.

To confirm the above observations statistically, two post-hoc Position (2) × Hierarchy (3) × Accent (2) three-way design ANOVAs were performed. In terms of syllable duration, two main effects, position and accent, are significant (Table 6.8). A two-way interaction effect involving position and accent is also significant. No effect involving hierarchy was found. The interaction effect comes from the fact that initial lengthening is evident only when syllables are accented, which makes syllable duration at P_0 and P_1 comparable [$t(166.61) = -.53, p > .5$].[1] When syllables are unaccented, duration at P_0 is significantly longer than that at P_1 [$t(197) = 5.76, p < .01$]. In other words, when syllables in both positions are pitch accented, there is no difference in duration between

Source	*df*	*F*	$\hat{\eta}^2$	*p*
		Between subjects		
Position (P)	1	4.59*	.01	.03
Hierarchy (H)	2	1.73	.01	.18
Accent (A)	1	143.17**	.27	.00
P × H	2	1.22	.01	.30
H × A	2	2.20	.01	.11
A × P	1	10.50**	.03	.00
P × H × A	2	1.03	.01	.36
Error	398	(0.03)		

*$p < .05$. **$p < .01$.

Table 6.8: Post-hoc ANOVA results for syllable duration in English with regards to accent patterning.

[1] Levene's Test for Equality of Variance showed that sphericity was violated ($p < .01$), therefore, adjusted *df* values were used. This is true throughout the study.

the two due to the lengthening effects at P_0 and P_1. However, when syllables in the two positions are not pitch accented, no initial lengthening effect was shown. Only the final lengthening effect was observed.

A similar post-hoc three-way ANOVA was conducted on SOI. As shown in Table 6.9, all three main effects are significant. A two-way interaction effect involving position and accent is also significant. Unlike syllable duration, the interaction effect comes from the fact that for SOIs at P_0, duration increase due to pitch accent is smaller than that at P_1 [P_0: $t(203) = -2.02, p < .05$; P_1: $t(124.86) = -7.68, p < .01$].

Since results in Chapter 4 regarding break indices showed that there is a relatively high incidence of BI3 (i.e., intermediate phrase boundary) and BI4 (i.e., intonation phrase boundary) at P_1 (see Figure 4.1), and BI3 or BI4 after the first syllable following the

Source	*df*	*F*	$\hat{\eta}^2$	*p*
		Between subjects		
Position (P)	1	159.14**	.29	.00
Hierarchy (H)	2	5.07**	.03	.01
Accent (A)	1	33.08**	.08	.00
P × H	2	0.91	.01	.40
H × A	2	2.43	.01	.09
A × P	1	6.69*	.02	.01
P × H × A	2	0.08	.00	.93
Error	398	(0.21)		

*$p < .05$. **$p < .01$.

Table 6.9: Post-hoc ANOVA results for SOI duration in English with regards to accent patterning.

boundary necessarily means accent (and nuclear accent), the initial lengthening effect found here might actually be a byproduct of final lengthening at the prosodic level, and is not related to the structural boundary in any sense. To test this hypothesis, a *t*-test on syllable duration at P_1 that are not of a BI3 or BI4 prosodic break was run using accent as a grouping factor. Results showed that accent is still a significant effect in determining syllable duration at this position [$t(142.15) = -10.36$, $p < .01$]. Accented syllables are on average 402 ms long while unaccented syllables are 160 ms long. This is also true of SOI [$t(118.31) = -8.49$, $p < .01$]. Accented SOIs are on average 451 ms long and unaccented SOIs are 170 ms long.

Taken together, this indicates that the lengthening effect for accented syllables at P_0 is a summation of the effects of final lengthening and pitch accents while lengthening of accented syllables at P_1 is solely due to pitch accents. Pitch accents also act as an overall lengthening factor for SOIs. However, there seems to be a ceiling effect for how much an SOI can be lengthened. At P_0, the duration increase due to accents is reduced because there is already a large amount of lengthening due to final lengthening whereas at P_1, this is not the case.

The statistical analyses confirmed the above observations concerning Figures 6.3 and 6.4. Structural boundaries are indicated by initial and final syllable and SOI lengthening. However, the two effects have different sources. Final lengthening is mainly due to the final lengthening effect while initial lengthening is incurred by pitch accents. The two lengthening effects have different scopes. The lengthening effect incurred by pitch accents is more focused on syllable than on SOI lengthening, while that incurred by

boundary is more effective on SOI than on syllable lengthening. Discourse hierarchy is indicated by degree of final SOI lengthening and initial syllable and SOI lengthening.

6.6 Guoyu Results

Figure 6.5 shows the syllable and pause duration patterning at various discourse boundaries in Guoyu. The bars in the upper and the lower parts of the graph indicate syllable and pause duration, respectively. The axes are the same as those in Figure 6.2.

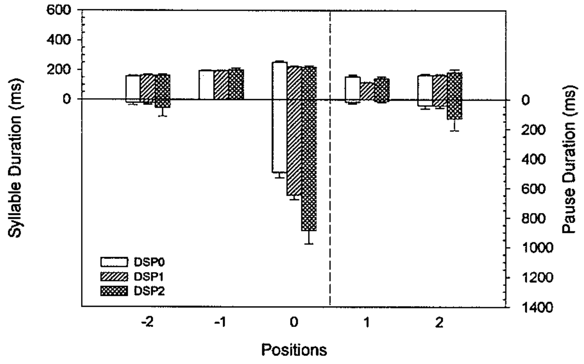

Figure 6.5: A bar graph illustrating the patterns of syllable and pause duration at structural boundaries in Guoyu. The *x*- and the *y*-axes are the same as those in Figure 6.3. The dashed line depicts the location of the boundary. The error bars indicate standard error.

From the figure, one notices that, unlike in English, the final lengthening effect at P_0 in Guoyu only stretches syllable duration slightly, although SOIs are still lengthened to a comparable extent. Another observation is that there is no duration increase of syllable that corresponds to the magnitude of discourse disjuncture. Instead, the degree of syllable lengthening tends to *decline* as discourse disjuncture becomes bigger, which is somewhat counterintuitive. However, boundary SOIs mirror discourse boundary strength in a positive manner as in English.

Two post-hoc Position (5) × Hierarchy (3) two-way mixed design ANOVAs were performed to confirm the above observations. For syllable duration, both of the main effects and the interaction effect are significant, as shown in Table 6.10. Post-hoc analyses horizontally across different positions using Bonferroni's adjustments showed that at DSP1 and DSP0 levels, syllables at P_0 and P_{-1} are the longest ($p < .01$) and the second longest ($p < .05$), respectively. At the DSP2 level, there is no significant difference between syllable duration at P_0 and P_{-1} ($p > .05$), although syllables at these two positions are still longer than those at other positions ($p < .01$). On the other hand, post-hoc analyses concerning the hierarchy effect using the Tukey's-*b* test showed that syllables at P_0 reflects boundary strength. Those at the DSP0 level are significantly longer than those at DSP1 and DSP2 levels ($p < .05$).

Source	*df*	*F*	$\hat{\eta}^2$	*p*
		Between subjects		
Hierarchy (H)	2	3.43*	.02	.03
Error	312	(0.01)		
		Within subjects		
Position (P)	3.68	58.62**	.16	.00
P × H	7.36	3.28**	.02	.00
Error	1147.69	(0.01)		

Note. Values enclosed in parentheses represent mean square errors. Due to violation of sphericity assumption using Mauchly's W test ($p < .01$), the *df*'s in the within-subjects section were adjusted using Huynh-Feldt adjustments.

*$p < .05$. **$p < .01$.

Table 6.10: Post-hoc ANOVA results for syllable duration in Guoyu.

A similar post-hoc two-way ANOVA concerning SOI was also performed. As shown in Table 6.11, both of the main effects and the interaction effect are again significant. Post-hoc analyses using Bonferroni's adjustments testing horizontally across the five positions showed that at all three DSP levels, SOIs at P_0 are the longest ($p < .01$). Post-hoc Tukey's-*b* tests also showed that hierarchically, duration of SOIs at P_0 mirrors boundary strength. SOIs at the DSP2 level are significantly longer than those at other levels ($p < .05$).

The statistical analyses confirmed the above observations. Both position and hierarchy effects are significant. But unlike in English, the position effect is observed at both P_0 and P_{-1} while the hierarchy effect is observed at P_0 only. Also, instead of the expected positive correlation between the amount of the lengthening effect and the degree of discourse disjuncture, the two are if anything negatively correlated. The relationship between discourse disjuncture and acoustic strength is nevertheless positively reflected in SOIs as in English.

Source	*df*	*F*	$\hat{\eta}^2$	*p*
		Between subjects		
Hierarchy (H)	2	11.32**	.07	.00
Error	312	(0.08)		
		Within subjects		
Position (P)	2.50	341.09**	.52	.00
P × H	5.00	5.56**	.03	.00
Error	779.64	(0.30)		

Note. Values enclosed in parentheses represent mean square errors. Due to violation of sphericity assumption using Mauchly's W test ($p < .01$), the *df*'s in the within-subjects section were adjusted using Huynh-Feldt adjustments.

*$p < .05$. **$p < .01$.

Table 6.11: Post-hoc ANOVA results for SOI duration in Guoyu.

6.7 Putonghua Results

Figure 6.6 shows the syllable and pause duration patterning at various discourse boundaries in Putonghua. The layout of the graph is the same as Figures 6.2 and 6.4. From the figure, one sees that there is very little lengthening at P_0 for syllable duration as compared to the syllable before it. As shown in Table 6.12, the syllable lengthening effect is on average a mere 30–40 ms regardless of the size of discourse disjunctures. In terms of SOI, the effect is more localized on the boundary position. Discourse hierarchy is also reflected mainly by SOIs.

Two post-hoc Position (5) × Hierarchy (3) two-way mixed design ANOVAs were performed to confirm the above observations. For syllable duration, only the main effect

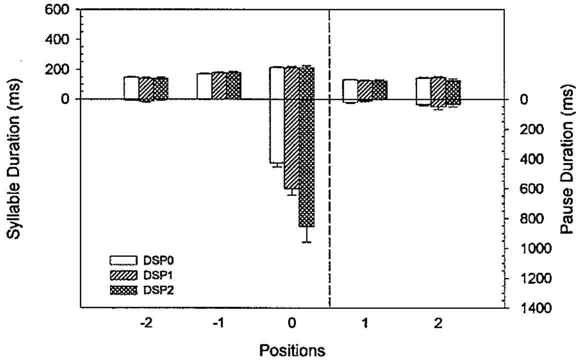

Figure 6.6: A bar graph illustrating the patterns of syllable and pause duration at structural boundaries in Putonghua. The *x*- and the *y*-axes are the same as those in Figure 6.3. The dashed line indicates the boundary location. The error bars indicate standard error.

(ms)	P_{-2}	P_{-1}	P_0	P_1	P_2
DSP0	146.22	169.93	211.44	131.19	141.73
DSP1	140.47	175.91	211.68	126.17	146.20
DSP2	141.04	177.43	213.27	123.42	125.19

Table 6.12: Average syllable duration (in ms) of the five positions around the boundary location in Putonghua.

Source	*df*	*F*	$\hat{\eta}^2$	*p*
		Between subjects		
Hierarchy (H)	2	0.24	.00	.79
Error	271	(0.01)		
		Within subjects		
Position (P)	3.61	53.72**	.16	.00
P × H	7.22	0.47	.00	.86
Error	978.09	(0.01)		

Note. Values enclosed in parentheses represent mean square errors. Due to violation of sphericity assumption using Mauchly's W test ($p < .01$), the *df*'s in the within-subjects section were adjusted using Huynh-Feldt adjustments.

*$p < .05$. **$p < .01$.

Table 6.13: Post-hoc ANOVA results for syllable duration in Putonghua.

of position is significant, as shown in Table 6.13. Post-hoc analyses using Bonferroni's adjustments showed that syllable duration is the longest at P_0 and the second longest at P_{-1} ($p < .01$).

A similar post-hoc two-way ANOVA concerning SOI was also performed. As shown in Table 6.14, both of the main effects and the interaction effect are significant. Post-hoc analyses using Bonferroni's adjustments testing horizontally across the five positions showed that at all three DSP levels, SOIs at P_0 are the longest ($p < .01$). Also, post-hoc Tukey's-*b* tests showed that duration of SOIs at P_0 mirrors boundary strength. Those at the DSP2 level are significantly longer than those at DSP1, and those at DSP1 are significantly longer than those at DSP0 ($p < .05$).

Source	*df*	*F*	$\hat{\eta}^2$	*p*
		Between subjects		
Hierarchy (H)	2	8.19**	.06	.00
Error	271	(0.06)		
		Within subjects		
Position (P)	1.64	356.94**	.57	.00
P × H	3.28	10.77**	.07	.00
Error	444.69	(0.13)		

Note. Values enclosed in parentheses represent mean square errors. Due to violation of sphericity assumption using Mauchly's W test ($p < .01$), the *df*'s in the within-subjects section were adjusted using Huynh-Feldt adjustments.

*$p < .05$. **$p < .01$.

Table 6.14: Post-hoc ANOVA results for SOI duration in Putonghua.

The statistical analyses confirmed the above observations of Figure 6.6. The final lengthening effect is at work at P_0 and P_{-1} for syllables but only at P_0 for SOIs. Discourse boundary strength is reflected only by SOIs.

6.8 Japanese Results

Figure 6.7 shows the syllable and pause duration patterning at various discourse boundaries in Japanese. The layout of the graph is the same as that for the other languages. Again, one sees that the syllable and SOI lengthening effects are localized at P_0. As in Guoyu, discourse hierarchy is reflected by a decrease in syllable duration and an

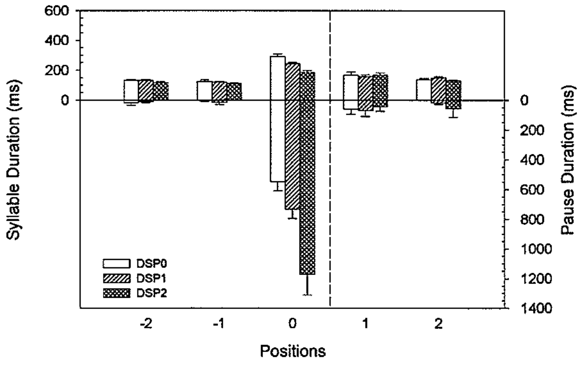

Figure 6.7: A bar graph illustrating the patterns of syllable and pause duration at structural boundaries in Japanese. The *x*- and the *y*-axes are the same as those in Figure 6.3. The dashed line indicates the boundary location. The error bars indicate standard error.

increase in SOIs at the boundary position. In other words, bigger disjunctures will result in shorter boundary syllables and longer boundary SOIs.

Two post-hoc Position (5) × Hierarchy (3) two-way mixed design ANOVAs were performed to confirm the above observations. For syllable duration, both of the main effects and the interaction effect are significant, as shown in Table 6.15. Post-hoc analyses using Bonferroni's adjustments showed that horizontally across different positions, syllable duration is the longest at P_0 at the DSP0 and DSP1 levels ($p < .01$). At the DSP2 level, however, there is no difference between syllable duration at P_0 and P_1, although syllables at P_0 are still longer than those at the other three positions ($p < .01$). Hierarchically, post-hoc Tukey's-*b* tests showed that syllable duration at P_0 corresponds to discourse boundary strength in a negative fashion. Syllable duration is the longest at the DSP0 level and the shortest at the DSP2 level ($p < .05$).

Source	*df*	*F*	η^2	*p*
		Between subjects		
Hierarchy (H)	2	5.38**	.07	.01
Error	135	(0.01)		
		Within subjects		
Position (P)	3.22	55.72**	.29	.00
P × H	6.45	3.38**	.05	.00
Error	435.22	(0.01)		

Note. Values enclosed in parentheses represent mean square errors. Due to violation of sphericity assumption using Mauchly's W test ($p < .01$), the *df*'s in the within-subjects section were adjusted using Huynh-Feldt adjustments.

**p* < .05. ***p* < .01.

Table 6.15: Post-hoc ANOVA results for syllable duration in Japanese.

A similar post-hoc two-way ANOVA concerning SOI was also performed. As in the ANOVA for syllable duration, both of the main effects and the interaction effect are significant (Table 6.16). Post-hoc analyses using Bonferroni's adjustments testing horizontally across the five positions showed that at all three DSP levels, SOIs at P_0 are the longest ($p < .01$). Hierarchically, post-hoc Tukey's-*b* tests showed that duration of

Source	*df*	*F*	$\hat{\eta}^2$	*p*
		Between subjects		
Hierarchy (H)	2	5.01**	.07	.01
Error	135	(0.09)		
		Within subjects		
Position (P)	1.92	196.42**	.59	.00
P × H	3.84	5.16**	.07	.00
Error	259.01	(0.21)		

Note. Values enclosed in parentheses represent mean square errors. Due to violation of sphericity assumption using Mauchly's W test ($p < .01$), the *df*'s in the within-subjects section were adjusted using Huynh-Feldt adjustments.

*$p < .05$. **$p < .01$.

Table 6.16: Post-hoc ANOVA results for SOI duration in Japanese.

SOIs at P_0 mirrors boundary strength. Those at the DSP2 level are significantly longer than those at DSP1 and DSP0 levels ($p < .05$).

The statistical analyses confirmed the above observations. The final lengthening effect is localized at P_0. At this position, the degree of syllable lengthening decreases when discourse boundary strength increases. On the other hand, SOI lengthening is reflective of discourse disjuncture in a positive manner.

6.9 Cross-linguistic Comparisons

Comparing across different languages, one sees that not only do the exact location and degree of localization of the final lengthening effect vary, the absolute value of

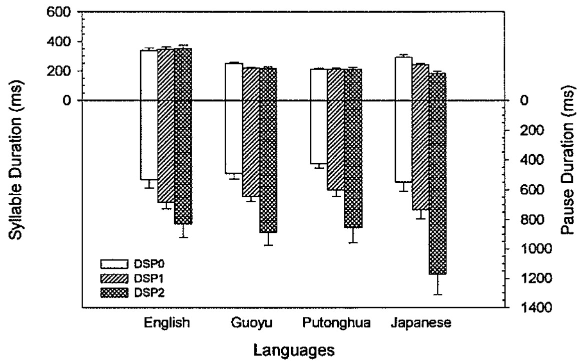

Figure 6.8: A bar graph illustrating the patterns of syllable and pause duration for boundary syllables (P_0) in all four languages. The *x*- and the *y*-axes are the same as those in Figure 6.3. The error bars indicate standard error.

lengthening and its relationship to the discourse hierarchy also differ from language to language. Figure 6.8 is a direct comparison of the boundary syllables and following pauses in the four languages. One notices that the magnitude of syllable lengthening is on average larger in English and Japanese and smaller in the two Mandarin varieties. On the other hand, the magnitude of pause lengthening, and thus SOI lengthening in Japanese is the largest of all four languages.

Post-hoc Tukey's-*b* tests showed that in terms of syllable duration, English and Japanese have significantly bigger lengthening effects at the DSP0 level than the two

Mandarin varieties ($p < .05$). At Levels DSP1 and DSP2, English has the longest boundary syllables ($p < .05$) while the other three languages do not differ much from one another. In terms of SOIs, however, the four languages have similar degrees of lengthening at all three levels, although boundary SOIs in English and Japanese are still slightly longer than Guoyu and Putonghua ($p < .05$).

6.10 Discussion

Since the first reports of final lengthening (Goldman-Eisler, 1972; Huggins, 1974; O'Malley, Kloker, & Dara-Abrams, 1973), the effect has been considered so ubiquitous that one tends to focus on its universal rather than language-specific aspects. However, the effect does seem to vary across languages, as shown in Sections 6.5–6.8. In English and Japanese, the lengthening effect is fairly localized at the boundary syllable P_0, as shown in Figures 6.3 and 6.7. All the other syllables prior to the boundary syllable and following the postboundary pause are more or less isochronous. On the other hand, in Putonghua and Guoyu, the two Mandarin varieties tested in this study, the final lengthening effect starts earlier, at around P_{-1} and peaks at P_0.

Also, just as a typical syllable denotes different durations in different languages (Figure 6.1), a typical lengthened syllable does not have the same duration across languages (Figure 6.8). The lengthening effect acts upon languages differentially in that boundary syllables in English and Japanese are stretched longer than those in the two Mandarin varieties. This is interesting in that it shows a typological split between English and Japanese on the one hand, and the two Mandarins on the other. English and Japanese

show a greater degree of lengthening that is more localized at the boundary syllable itself while Guoyu and Putonghua show a smaller degree of lengthening, yet their lengthening effect is more spread out.

The fact that Japanese has a greater lengthening effect on boundary syllables is also interesting. On average, Japanese has the shortest mean and median syllable durations of all four languages (see Figure 6.1). This implies that syllable stretchability is also language-specific. Of the four languages studied here, Japanese syllable is the most stretchable.

The interaction between syllable duration and pitch accents in English is an intriguing phenomenon, too. Although accented syllables tend to be longer in general, they are especially lengthened at DSP-initial positions (i.e., P_1, see Figure 6.4). SOIs, on the other hand, are not affected by pitch accents as much. In other words, the lengthening effect of pitch accents is more localized on syllable rather than on pause, and more on P_1 than any other positions examined.

In contrast, the lengthening effect in SOIs is more universal. For all four languages, the effect is focused on the pause at P_0. Not only that, the magnitude of the lengthening effect is about the same across languages (Figure 6.8). It seems that there is a universal optimal range as to how long an SOI can be. It does not matter as much the proportions of the compositions in an SOI. Syllable and pause duration are in a compensatory situation. When a syllable is shorter (as in Japanese), pause becomes longer.

The lengthening effect of syllable duration is not only reflective of the location of a structural boundary, but it is also in some form indicative of boundary strengths in all four languages studied here. There are three recognized levels of discourse disjuncture in this study—DSP0, DSP1, and DSP2. DSP0 represents a smooth transition between two adjacent clauses (i.e., without a discourse disjuncture), DSP1 a low degree of discourse disjuncture, and DSP2 a high degree of discourse disjuncture. In terms of acoustic measurements, English, Guoyu, and Japanese all displayed a trend that is reflective of this hierarchy. However, statistically speaking, English only distinguishes the highest level of disjuncture (i.e., DSP2 vs. DSP1 & DSP0), Guoyu makes a distinction between clausal and discourse disjuncture (i.e., DSP0 vs. DSP1 & DSP2), while Japanese distinguishes all three levels (i.e., DSP0 vs. DSP1 vs. DSP2). Surprisingly though, the relationship between boundary and acoustic strengths is not necessarily a positive one, as one would naïvely predict. Instead, the degree of lengthening can decrease as boundary strength increases. What is more surprising is that this is not a single case, as one finds that both Guoyu and Japanese demonstrated this pattern (Figures 6.5 & 6.7). English, on the other hand, showed the expected pattern of positive correlation at P_0 in addition to the initial strengthening effect found with accented syllables at P_1. Putonghua does show any correlation between discourse hierarchy and the magnitude of the syllable lengthening effect.

On the other hand, SOI patterning in all four languages is reflective of discourse hierarchy. Statistically speaking, however, English and Guoyu only distinguish the high level of disjuncture from the other two (i.e., DSP2 vs. DSP1 and DSP0) while Putonghua and Japanese make a distinction with all three levels (i.e., DSP0 vs. DSP1 vs. DSP2). The relationship between SOIs and discourse hierarchy is also more universal in that at each level of discourse disjuncture, there is about the same magnitude of lengthening across all

four languages (Figure 6.8). Not only that, SOIs always correlate with discourse boundary strength in a positive manner. Since SOIs are composed of syllables and the following silent pauses, this implies that pause duration reflects disjunctures positively. If one looks at the patterning of syllable and pause duration together, it is not unreasonable to assume that the cross-linguistically different surface strategies with syllable lengthening might actually stem from a more universal goal, that is, to lengthen the absolute and relative duration of the following silent pause as discourse disjuncture increases. This would explain why Guoyu and Japanese adopt a seemingly unintuitive strategy by decreasing the amount of lengthening to indicate bigger discourse disjunctures. Of course, this does not mean languages have to lengthen pause duration at the expense of syllable duration. English and Putonghua are two counterexamples. This would just imply that languages do have different preferences as to how to indicate discourse hierarchy, and it just so happens that Guoyu and Japanese emphasize more on pause duration than English and Putonghua.

CHAPTER 7

GENERAL DISCUSSION

This study looked at the distribution of categorical prosodic boundaries and variation in the acoustic measures relative to sentential boundaries within and across discourse segment boundaries. Boundary and hierarchy effects were found but they were not active to the same extent across parameters and across languages. In the following, boundary and hierarchy effects and their implications are discussed. How these effects interact with language-specificity is also considered.

7.1 Boundary Effect

Acoustically and prosodically, to have a boundary effect with regards to sentential and discourse boundaries is not surprising. Historically, studies have shown that such an effect is manifested to various degrees (Cooper & Sorensen, 1977; Fon & Johnson, 2000; Goldman-Eisler, 1972; Grosz & Hirschberg, 1992; Hirschberg & Nakatani, 1996; Koopmans-van Beinum & Donzel, 1996; Ladd, 1988; Lehiste, 1975a, 1975b, 1979a, 1979b; Lehiste & Wang, 1977; Oller, 1973; Swerts, 1998; Swerts & Geluykens, 1994;

Terken, 1984; Traum & Heeman, 1997). Sentential boundary cues are also prevalent in child-directed speech, often with an exaggerated magnitude, which presumably helps language acquisition (Fisher & Tokura, 1996; Ratner, 1986). Moreover, listeners are able to detect such signals (Berkovits, 1984; Kreiman, 1982; Price, Ostendorf, Shattuck-Hufnagel, & Fong, 1991; Sanderman & Collier, 1996; Shen, 1992; Speer, Kjelgaard, & Dobroth. 1996; Swerts, 1997; Swerts & Geluykens, 1994), although whether such signals can be effectively used in online segmentation is still of some debate (Schaffer, 1984). Studies have shown that infants either are predisposed to such signals or develop abilities to detect them from a very early stage (Kemler Nelson, Hirsh-Pasek, Jusczyk, & Cassidy, 1989).

However, not all studies dealt with spontaneous speech, and genre differences might contribute to some of the discrepancies in the results. Also, when looking at sentential cues, many studies focus on ambiguous sentences rather than sentences without ambiguity (Price, Ostendorf, Shattuck-Hufnagel, & Fong, 1991; Speer, Kjelgaard, & Dobroth, 1996; Carlson, Clifton, & Frazier, 2001). It is unclear whether the cues examined still exist when the speaker is not aware of an ambiguity (Lehiste, Olive, & Streeter, 1976). Finally, except for a couple (Dutch: Koopmans-van Beinum & Donzel, 1996; Swerts, 1997, 1998; Swerts & Geluykens, 1994; Terken, 1984; Japanese: Fisher & Tokura, 1996; Mandarin: Fon & Johnson, 2000), most studies concentrated and thus based their frameworks only on the results of English. Thus, the significance of the results in this study lies in its spontaneity, its non-ambiguity, and its cross-linguistic comparisons.

ToBI break indices were used to measure prosodic breaks. In general, there is not much cross-linguistic difference in prosodic break patterning with regards to structural boundary cues, which is what one would expect, if prosodic boundaries are categorical. Using the percentage of intonation phrase (IP) breaks as the dependent variable, one finds there to be a universal trend for IP breaks to occur more frequently at the boundary syllable than at other non-boundary positions. However, there are proportionately more IP breaks at the boundary in English and Japanese as compared to Guoyu and Putonghua. This is due to the observation that in Guoyu and Putonghua, break indices that are lower than IP breaks (i.e., from BI2 in M-ToBI and above) are allowed at the boundary position, which is usually not the case in English and Japanese.

The binary distinction between the two Mandarin varieties and English and Japanese might not be solely due to language differences. One other possible cause might be the immaturity of M-ToBI, which is a newly developed and still developing system that has not undergone extensive intertranscriber reliability tests. Also, since the current M-ToBI system is combined from systems developed in two sites, Academia Sinica in Taipei, Taiwan, and The Ohio State University in Columbus, Ohio, there might be some incongruence in the basic designs of the framework. More studies concerning the M-ToBI need to be done in order to test this hypothesis.

Turning to acoustic-phonetic measures, one finds pitch peak trend to be a more variable parameter as a boundary indicator across languages. In English, it does not show any consistent pattern at all. In all the other languages, there is some indication of pitch reset at a structural boundary. For the two varieties of Mandarin, pitch reset is only consistently found at the DSP1 and DSP2 levels, while for Japanese, it is found at all

three levels. A consistent declination trend within a structural unit was found in Japanese but not in Mandarin. This is probably due to the differences in tonal distribution in the two languages. Mandarin is more densely packed with tones than Japanese, and the trend for declination might be more easily revealed by a longer stretch of utterances.

However, the fact that pitch peak trend does not play a consistent role at boundary locations in English cannot be completely due to the sparseness of tones in the language since Japanese also has a sparse distribution of tones. It simply implies that the pitch peak trends are more variable in English than those in other languages. A preliminary look into the intonation patterns showed that the lack of declination and pitch reset might be due to the uptalk style that is prevalent among young central Ohio speakers. More studies are needed in order to answer this question.

Gender also plays a role in complicating the patterning of pitch peak trends. For English and Guoyu, female speakers have more drastic pitch trend manipulations than male speakers. For Putonghua and Japanese, it is the opposite. Male speakers on average have more obvious pitch trend manipulations than female speakers. Overall speaking, English and Putonghua female speakers, and Japanese and Putonghua male speakers have the highest pitch values within gender around the boundary region, while Guoyu and Japanese females, and English and Guoyu males have the lowest. As a consequence, gender difference regarding pitch value is the largest in English and smallest in Japanese.

Of the two acoustic-phonetic cues examined, duration seems to be the most universal in identifying clausal and discourse boundaries; all four languages examined make some use of duration cues. Final lengthening on boundary syllables and SOIs is consistently found in all four languages, although the magnitude differs across languages.

English and Japanese have the highest degree of boundary syllable lengthening as compared to the two Mandarins. This difference is mitigated when the duration of the following pause (i.e., SOI) is also taken into consideration.

The fact that Japanese has a high degree of syllable lengthening as compared to other languages is interesting since Japanese also has the shortest mean and median syllable duration. This is in conflict with what Campbell and Isard's (1991) model would predict. Their model had an invariant component, which was constant across different inherent durations, and a variable component, which was proportional to intrinsic duration. Following that, one would predict that languages with the shortest durations should show the least elasticity, and thus show the smallest amount of lengthening, which is not the case in this study.

The final lengthening effect seems to universally affect the boundary position (i.e., P_0). However, other positions at the boundary region also seem to be affected by some lengthening effect in some of the languages. In English, in addition to final lengthening, there is also an initial lengthening effect at utterance-initial positions. However, initial lengthening only occurs when the syllable/SOI is pitch-accented. This is true even when BIs of 3 and 4 (i.e., intermediate and intonation phrase boundaries) are excluded. The pattern is similar to what Fougeron and Keating (1997) have proposed in that articulatory strengthening occurs at both edges of a prosodic domain, although they did not test sentences without initial accents. The magnitude of the two effects, initial and final lengthening, is comparable in terms of syllable duration, but not SOI. For SOIs, final lengthening is more effective in lengthening the duration, even when accents are controlled for.

The locus for syllable lengthening is also widened in Guoyu and Putonghua, but in the other direction. Both the ultimate and the penultimate boundary syllables undergo final lengthening, although the former is still lengthened to a greater extent than the latter. However, this gradient effect of lengthening is not found in SOI.

Table 7.1 is a summary of the structural boundary cues examined in this study. IP boundary and final lengthening at boundary syllables and SOIs, and prosodic breaks of nonboundary syllables seem to be universal. Except for English, pitch reset also occurs consistently as a structural boundary cue. There are also cues that seem to be more language-specific. For example, final lengthening at P_{-1}, and prosodic breaks lower than an IP boundary at boundary syllables are specific to the two varieties of Mandarin. The weak final lengthening effect is also characteristic of this language. English is peculiar in that it has initial lengthening at P_1 when pitch-accented. In terms of pitch range, Putonghua and Japanese show a surprising pattern of male having a more drastic pitch peak trend than female. On the other hand, English and Japanese sometimes also pattern together, as in IP prosodic breaks at boundary positions. In addition, the two languages are more comparable in the magnitude of final lengthening.

7.2 Hierarchy Effect

That spontaneous speech is organized in a hierarchical fashion is not a new idea. However, most of the studies focus on the prosodic organization of speech when talking about hierarchy (Beckman, 1996; Ferreira, 1993; Selkirk, 1984, 1986). Although one cannot deny that prosody has its own hierarchical organization, there is also evidence

	English	Guoyu	Putonghua	Japanese
Prosodic breaks				
Boundary syllable	IP	≤ IP	≤ IP	IP
Nonboundary syllable	< IP	< IP	< IP	< IP
Pitch peak trend				
Pitch reset	No	Yes	Yes	Yes
Gender	M < F	M < F	M > F	M > F
Duration				
Syllable				
Final lengthening at P_0	YES	Yes	Yes	YES
Final lengthening at P_{-1}	No	Yes	Yes	No
Initial lengthening at P_1	Yes	No	No	No
SOI				
Final lengthening at P_0	YES	Yes	Yes	YES
Initial lengthening at P_1	Yes	No	No	No

Table 7.1: A summary of the prosodic and acoustic-phonetic cues for the four languages. Capitalization indicates a bigger magnitude as compared across languages. 'M' and 'F' stand for male and female, respectively.

showing that structural hierarchy can be reflected by prosodic and acoustic measures. The hierarchy revealed can be as small as at a phrasal (Huggins, 1974), a clausal (Ladd, 1988), or a discourse level (Fon & Johnson, 2000; Grosz & Hirschberg, 1992; Hirschberg & Nakatani, 1996).

In this study, both prosodic and acoustic-phonetic parameters are examined to see whether discourse hierarchy is encoded in speech. Prosodic breaks are reflective of hierarchy in some sense. Using percentage of IP breaks as a dependent measure, one finds that the proportion of IP breaks at P_0 is indicative of hierarchy in all four languages. Proportionately more IP boundaries occur at higher discourse disjunctures. In English, the proportion of IP breaks at P_1 is also reflective of hierarchy. Of the four languages examined, the hierarchy effect reflected by percentage of IP breaks is stronger for the two Mandarins and weaker for English and Japanese. This is due to the fact that in Guoyu and Putonghua, low level prosodic boundary breaks are allowed at the boundary position for lower discourse disjunctures, which is not the case in English and Japanese. The two languages almost always allow only IP breaks at the boundary position regardless of level of disjuncture.

Pitch reset is only a possible hierarchy cue in Guoyu, Putonghua, and Japanese. English, on the other hand, does not have a consistent patterning of this parameter with regard to boundary. For the Putonghua, there is a positive correlation between the amount of pitch reset and declination, and hierarchy at the boundary region. For Guoyu, this is only true for female speakers. For male speakers, there is only a correlation between the amount of pitch reset and hierarchy. For Japanese, male speakers show a positive correlation between the amount of reset and hierarchy, while female speakers show a positive correlation between the degree of declination and hierarchy. The gender effect in Japanese might be due to a sociolinguistic factor. Traditionally, Japanese female speakers tend to use the higher portions of their pitch range while Japanese male speakers tend to use the lower. Therefore, it might be physiologically easier for Japanese male speakers to manipulate degree of reset and Japanese female speakers degree of declination at the boundary region. This might also explain why Guoyu male speakers use only pitch reset to reflect hierarchy, since according to Figure 5.6, Guoyu male speakers are lower in pitch values as compared to the other languages. At any rate, for the three languages that show a consistent pitch reset and declination, all of them make statistically significant distinctions only between a high level of discourse disjuncture and the lower levels (i.e., DSP2 vs. DSP1 & DSP0).

Like the boundary effect, duration seems to be the most universal and sensitive cue in this respect. All four languages show a positive correlation between SOI and hierarchy at the boundary position. English in addition also shows a reflection of hierarchy at P_1 when the syllable in the position is pitch accented. English, Guoyu, and Japanese make a distinction only between the higher level of disjuncture (i.e., DSP2) and the lower levels (i.e., DSP1 & DSP0), while Putonghua makes a distinction among all three levels.

Syllable duration is more variable in cueing hierarchy. Only Guoyu and Japanese use boundary syllable duration to cue hierarchy, and both in a negative fashion. That is, bigger sized disjunctures are encoded by shorter syllable duration. Guoyu only makes a distinction between clausal and discourse boundaries (i.e., DSP0 vs. DSP1 & DSP2) while Japanese distinguishes all three levels. Boundary syllable duration in English and

Putonghua is not reflective of structural hierarchy at all. However, when syllables are pitch accented at utterance-initial positions, syllable duration can also be reflective of hierarchy in English.

Table 7.2 is a summary of the hierarchy cues examined in this study. Percentages of IP breaks and Final SOI lengthening seem to be the most universal cue in this respect. All four languages make a distinction between the highest level of disjuncture (i.e.,

	English	Guoyu	Putonghua	Japanese
Prosodic breaks				
P_0	2 > 1, 0	2 > 1 > 0	2 > 1, 0	2, 1 > 0
P_1	2 > 1, 0	No	No	No
Pitch peak trend				
Pitch reset	---	2 > 1, 0	2 > 1, 0	2 > 1, 0 (M)
Declination	---	2, 1 > 0 (F)	2, 1 > 0	2, 1 > 0 (F)
Duration				
Syllable				
Final lengthening at P_0	No	0 > 1, 2	No	0 > 1 > 2
Initial strengthening at P_1	2 > 1, 0	---	---	---
SOI				
Final lengthening at P_0	2 > 1, 0	2 > 1, 0	2 > 1 > 0	2 > 1, 0
Initial strengthening at P_1	2 > 1, 0	---	---	---

Table 7.2: A summary of the acoustic and prosodic hierarchy cues for the four languages. The numbers ‘0’, ‘1’, and ‘2’ refer to the three disjuncture levels DSP0, DSP1, and DSP2, respectively. ‘M’ and ‘F’ refer to male and female, respectively. ‘---‘ indicates not applicable.

DSP2) and the lower levels (i.e., DSP1 & DSP0). Putonghua further distinguishes between DSP1 and DSP0. Pitch peak trend is more language-specific. In Putonghua, the degree of reset and declination reflect hierarchy regardless of gender, while in Guoyu, this is true only for female speakers. Male Guoyu speakers only use pitch reset to indicate hierarchy. In Japanese, there is a gender split. Male speakers use pitch reset, and female speakers declination to reflect hierarchy.

Relatively speaking, English is the most different of all four languages. First of all, initial lengthening on syllable and SOI duration reflects hierarchy. Also, patterning in pitch peak trend is not a robust indicator of hierarchy, and percentage of IP breaks reflect hierarchy at P_1 in addition to P_0.

Table 7.3 summarizes the levels of distinctions made by the four languages. As shown in Table 7.2 above, languages have various ways of making a distinction, through prosodic breaks, reset, declination, lengthening, and so forth. As long as languages make a consistent distinction of a certain level, there is a chance for listeners to take advantage of the encoding and use it as an aid for segmentation and acquisition. As shown in Table 7.3, only Guoyu, Putonghua, and Japanese encode all three levels. English only consistently encodes two levels of distinction—DSP2 versus DSP1 and DSP0. This implies that English only distinguishes a high degree of discourse disjuncture from other low-level disjunctures while for the other languages, the three levels are distinguished somehow to a certain extent.

Phenomena such as final lengthening and pitch reset at the structural boundary region have been assumed for decades (e.g., Cooper & Sorensen, 1977; Lehiste, 1975a; Streeter, 1978). However, when only one or a few languages are the focus of study, it is easy to fall a prey to overgeneralization. Take the most prevalent and supposedly the most universal final lengthening for example, there are still cross-linguistic differences regarding the locus of lengthening, and how the effect interact with accent and hierarchy. Other more variable phenomena such as pitch reset are even more language-dependent.

	English	Guoyu	Putonghua	Japanese
DSP2 vs. DSP1, DSP0	Yes	Yes	Yes	Yes
DSP2, DSP1 vs. DSP0	No	Yes	Yes	Yes
DSP2 vs. DSP1 vs. DSP0	No	Yes	Yes	Yes

Table 7.3: A summary of the distinctions made among the three hierarchy levels. 'Yes' refers to cases where there are cues that can significantly tell the levels apart. 'No' refers to cases where such a trend exists, but does not reach a significant level.

Figure 7.1 is a schematic summary of the boundary and hierarchy cues examined in this study. The horizontal line indicates a continuum of language universality. To the right of the graph are the cues that are more language-universal, and to the left of the graph are those that are more language-specific. The degree of universality is provisionally determined by looking at the four languages examined in this study. A cue is considered more universal if it is shared by more languages. The upper part of the figure is for boundary cues while the lower part of the graph is for hierarchy cues. One sees that boundary SOI lengthening is the most universal in terms of both the boundary and the hierarchy cues, while initial lengthening is the least universal. Also, boundary cues are in general more universal than hierarchy cues.

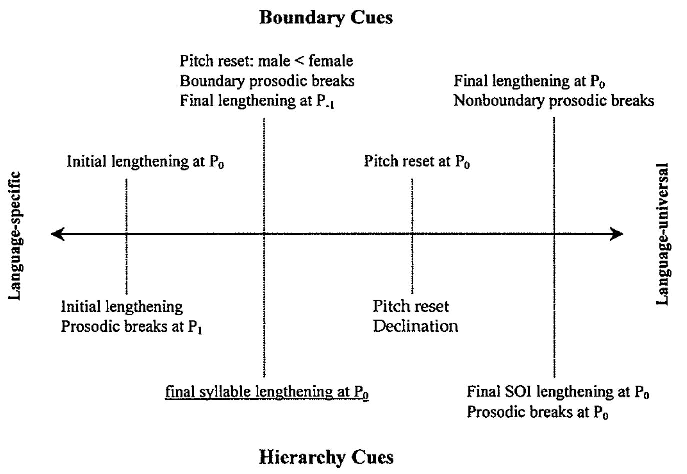
Boundary Cues
Pitch reset: male < female
Boundary prosodic breaks
Final lengthening at P_{-1}
Final lengthening at P_0
Nonboundary prosodic breaks
Initial lengthening at P_0
Pitch reset at P_0
Language-specific
Language-universal
Initial lengthening
Prosodic breaks at P_1
Pitch reset
Declination
final syllable lengthening at P_0
Final SOI lengthening at P_0
Prosodic breaks at P_0
Hierarchy Cues

Figure 7.1: A schematic figure showing the relative degree of language-universality of the cues examined in this study. The right hand side represents language-universal cues while the left hand side language-specific cues. The top part of the graph indicates the boundary cues while the bottom part of the graph indicates the hierarchy cues. In the bottom half, an underline indicates a negative correlation. All the others are positive.

From the figure, one sees that there is only one cue whose magnitude is in negative correlation with hierarchy, that is, final syllable lengthening at P_0 (indicated by an underline). Of the four languages examined, only Guoyu and Japanese are involved. All the other cues are in positive correlation with discourse boundary hierarchy.

Negative correlation is unintuitive at a first glance. However, it is not totally unconceivable since cues in spontaneous speech are gauged in context. By shortening the degree of final boundary syllable lengthening, the relative duration of the following pause is increased. Therefore, the diminished degree of lengthening is actually a byproduct of a lengthening effect on an adjacent element (or non-element, in this case). In other words, when looking at spontaneous speech, it is necessary to consider cues in context. Speech cues considered in isolation can lead to unintuitive and puzzling results. However, when various cues are considered in tandem, it is not difficult to realize that cues are working in a cooperative manner to aid online speech segmentation, regardless of the direction of correlation.

One also finds that cues are often patterned together in a similar fashion within a language as compared to across languages. This is evidenced by the fact that both varieties of Mandarin, Guoyu and Putonghua, show a very similar pattern in a number of parameters. For example, in terms of final syllable lengthening, both show an effect at P_{-1}. The two languages also have similar distributions concerning boundary breaks. Prosodic boundaries that are lower than an IP break can occur at the boundary position.

In terms of hierarchy cues, Japanese patterns more like the Mandarins, and English tends to be the peculiar one. For example, English has initial lengthening reflecting hierarchy, which is not found in all the other languages. Also, pitch reset is not reflective of hierarchy in English while in all the other languages, pitch reset is a fairly robust cue. This is an interesting finding since as mentioned above, most of the studies are based on English data. However, as it turned out in this study, the patterning in English is not the majority. It is actually fairly different from the other languages. This is by no means to say that what is found as the majority patterning here is the majority in reality, although such a possibility does exist. Instead, what should be kept in mind is that one's belief can be easily based on a biased and convenient sampling. True universality can only be approximated through extensive cross-linguistic studies.

CHAPTER 8

CONCLUSION

This study has aimed itself to look at three effects: boundary, hierarchy, and language. Results are promising. Structural boundaries are signaled by a combination of cues—IP boundary breaks, pitch reset, and initial and final lengthening. The realization of effects is to a certain extent determined by individual languages. Languages can have somewhat peculiar ways of indicating structural boundary. For example, English uses initial lengthening in addition to final when pitch-accented syllables are encountered. No consistent usage of pitch reset was found in this language. Mandarin shows early and only a modest degree of final lengthening. In addition, prosodic breaks at a level lower than an IP break are allowed at boundary positions. Japanese patterns like Mandarin in that it also makes use of pitch reset for boundary cues. However, it also patterns like English in that only prosodic breaks of IP boundary level or higher are found at sentential/discourse boundaries.

Hierarchy is often signaled by modulations of boundary cues. Thus, theoretically, all of the boundary cues are potential candidates. As lengthening is the most robust boundary cue, it is not surprising to find that SOI lengthening is also the most consistent

cue for hierarchy. However, individual languages also have their ways of painting their own pictures. English, for example, takes advantage of the initial lengthening phenomenon for accented syllables and makes the cue gradient to show hierarchy. This is still true when syllables at P_1 are controlled for break indices. Following the tradition of focusing on the utterance-initial rather than the utterance-final position, utterance-initial IP break percentages in English also modulate with hierarchy. At this position, more IP breaks are found at higher structural hierarchy.

Although Guoyu and Putonghua belong to the same language, they are different in certain respects. In Putonghua, both genders use the degree of pitch reset and declination to indicate hierarchy. In Guoyu, this is only true for female speakers. For male speakers, only the degree of pitch reset is used. The two languages also differ in their usages of final syllable lengthening. Guoyu speakers use it as a gradient indicator for hierarchy in Guoyu but not in Putonghua.

As for Japanese, even though it is linguistically not related to Mandarin, it sides with the two Mandarins more than it does with English. For example, it patterns more like Guoyu in terms of syllable final lengthening and pitch reset and declination, although its IP break distribution is more like English. However, Japanese also has its peculiarity. For example, in the pitch domain, male speakers only use pitch reset as a hierarchy indicator, while female speakers only use declination.

It is exciting to find cross-linguistic differences as a cross-linguistic study. However, the aims for such a study do not lie merely in the actual patterning. Instead, what one attempts to achieve in a study like this is to have a more thorough understanding of the phenomenon with a broad range of languages. By doing so, one hopes to provide some insight into how speech is organized hierarchically in a universal yet also specific way. This would be important when considering frameworks for segmentation in speech processing and language acquisition.

Naturally, cross-linguistic studies do not stop at only four languages, and segmentation cues in spontaneous speech cannot possibly be covered only by 32 around-five-minute narrations. However, one would like to deem this as a promising start. It is hoped that more studies on other languages and also other genres will be possible in the future. One would also like to turn to the listeners' side and see whether the cues found can also be perceived and utilized at some unconscious level of processing. This would be especially interesting and relevant to issues in language acquisition and language universality since as shown in this study, one already found sets of cues that are more universal (within the four languages) than others.

It is not hard to imagine perceptibility of native cues for boundary and hierarchy. However, how foreign cues are perceived, or even, whether they can be perceived is a fundamental issue for (second) language acquisition. There might be cue ordering and cue weighting in that some cues are easier to perceive than others. In addition, certain cue patterning might facilitate acquisition more while others might prove to be difficult. The degree of ease of acquisition might also be directly related to similarity in cue packaging between the native language and the language to be acquired. All these questions definitely merit long-term future research.

It is hoped that by providing a simple glimpse of the phenomenon, one can obtain a better understanding of how language is and what language is. It is also hoped that as more and more cross-linguistic and non-English studies are made available, a language

processing framework that is powerful enough to depict language-universals, yet also generous enough to accommodate language-specifics can be developed eventually.

APPENDIX A

DEMOGRAPHICS OF RECRUITED SUBJECTS

Tables A.1 through A.4 show information concerning the subjects that participated in this study. The languages concerned are English, Guoyu, Putonghua, and Japanese, accordingly.

ID	Sex	Age	Race	Childhood[a]	Education	Language Ability
BH	male	40	Caucasian	Columbus, OH	College	English (0)
JH	male	26	Caucasian	Columbus, OH	Post-grad	English (0), French (20)
MK	male	21	Caucasian	Urbana, OH Columbus, OH	College	English (0), Spanish (14)
PA	male	23	Caucasian	London, OH Columbus, OH	College	English (0)
AL	female	24	Caucasian	Columbus, OH	Post-grad	English (0), Portuguese (18), French (20), Spanish (22/23)
DF	female	29	Caucasian	Columbus, OH	College	English (0), German (12), French (18)
MD	female	19	Caucasian	Columbus, OH	College	English (0), Spanish (14)
SU	female	20	Indian	Dublin, OH	College	English (0), French (13)

[a] Childhood years include the years between ages 3 and 18.

Table A.1: Demographic information concerning English subjects. The numbers in the parentheses indicate the age one starts learning the language.

ID	Sex	Age	Race[a]	Childhood[b]	Education	Language Ability
CHL	male	30	Mainlander	Taipei	Post-grad	Guoyu (0), Shanghaihua (0), English (12), Taiwanese (12)
HPH	male	29	Taiwanese	Taipei	Post-grad	Guoyu (0), Taiwanese (?), English (13)
SCH	male	30	Taiwanese	Taipei	Post-grad	Guoyu (0), Taiwanese (0), English (12)
SCY	male	26	Taiwanese	Taipei	College	Guoyu (0), Taiwanese (5), English (12), Japanese (15)
CYW	female	28	Taiwanese	Taipei	Post-grad	Guoyu (0), Taiwanese (0), English (12)
HHY	female	29	Taiwanese	Sanchung, Chungho	Post-grad	Taiwanese (0), Guoyu (6), English (13)
SMI	female	26	Taiwanese	Taipei	Post-grad	Guoyu (0), Taiwanese (0), English (11)
WCF	female	28	Taiwanese	Panchiao, Taipei	Post-grad	Guoyu (0), English (10), Taiwanese (20), Japanese (20)

[a]Race is defined linguistically rather than ethnically here. Mainlanders usually grow up in a language environment that is mainly Guoyu and sometimes also some other non-Taiwanese, non-Hakka Chinese languages. On the other hand, Taiwanese usually grow up in a mixed language environment of Guoyu and Taiwanese, even if they do not speak Taiwanese themselves.

[b]Childhood years include the years between ages 1 and 22. Sanchung, Chungho, and Panchiao are cities within Taipei county.

Table A.2: Demographic information concerning Guoyu subjects. The numbers in the parentheses indicate the age one starts learning the language. '?' denotes not sure.

ID	Sex	Age	Race[a]	Childhood[b]	Education	Language Ability
FBS	male	37	Beijing	Beijing	Post-grad	Putonghua (0), English (18), Japanese (27)
LJG	male	29	Beijing	Beijing	Post-grad	Putonghua (0), English (9)
XD	male	33	Anhui	Beijing	Post-grad	Anhuihua (0), Putonghua (0), Pekingese (6)
YYL	male	28	Hebei	Beijing	Post-grad	Putonghua (0), English (12)
CY	female	26	Beijing	Beijing	Post-grad	Putonghua (0), English (12), German (21)
LX	female	27	Liaoning	Beijing	College	Putonghua (0), English (12)
ZLE	female	28	Beijing	Beijing	College	Putonghua (0), English (13), Italian (17)
ZLI	female	25	Beijing	Anhui,[c] Beijing	Post-grad	Putonghua (0), English (12)

[a]Race is defined linguistically rather ethnically here. This is formally defined as the place where one's father was born. All three providences here, Anhui, Hebei, and Liaoning belong to the Mandarin language area. Anhui belongs to the eastern/southern dialect while Hebei and Liaoning belong to the northern dialect.

[b]Childhood years include the years between ages 1½ and 18.

[c]Although ZLI spent her first eight years of life in Anhui, an eastern providence in China, her life was limited to a research institution in Hefei, the capital of Anhui because of her father's job as a researcher. According to her and the native speaker recording the corpus, such research institutions are usually situated on a very large piece of land and are fairly self-sufficient. They tend to form language islands of Beijing Putonghua since researchers and their families are somewhat restricted to the institute area.

Table A.3: Demographic information concerning Putonghua subjects. The numbers in the parentheses indicate the age one starts learning the language.

ID	Sex	Age	Birthplace[a]	Childhood[b]	Education	Language Ability
FY	male	30	Wakou	Wakou	Post-grad	Japanese (0), English (10)
IA	male	28	Tokyo	Inagi[c]	Post-grad	Japanese (0), English (12)
MS	male	31	Tokyo	Tokyo	Post-grad	Japanese (0), English (12)
NT	male	30	Odawara	Hong Kong,[d] Odawara	Post-grad	Japanese (0), English (4/12), Spanish (18)
NY	female	20	Shyouwa	Shyouwa	College	Japanese (0), English, (12), French (19)
SA	female	20	Yokohama	Yokohama	College	Japanese (0), English (12), German (19), Italian (19)
UK	female	19	Tokyo	Tokyo	College	Japanese (0), English (12)
YE	female	34	Tokyo	Tokyo	Post-grad	Japanese (0), English (12)

[a] Wakou and Shyouwa belong to Saitama Prefecture; Odawara and Yokohama belong to Kanagawa Prefecture.

[b] Childhood years include the years between ages 2 and 15.

[c] Inagi is within Tokyo Prefecture.

[d] NT lived in Hong Kong from ages 4 to 8 due to a position transfer of his father. NT went to a British kindergarten, and then went to a Japanese elementary school there. There was contact with local Cantonese-speaking children, but both the subject and his mother claimed that Cantonese knowledge was virtually gone a few months after their return to Japan.

Table A.4: Demographic information concerning Japanese subjects. The numbers in the parentheses indicate the age one starts learning the language.

APPENDIX B

RECORDING DURATION OF THE SUBJECTS

English		Guoyu		Putonghua		Japanese	
ID	duration	ID	duration	ID	duration	ID	duration
BH	3:19:522	CHL	2:00:471	FBS	2:22:028	FY	3:14:726
JH	3:32:414	HPH	9:19:123	LJG	2:22:327	IA	2:59:794
MK	2:10:477	SCH	2:19:700	XD	2:35:572	MS	4:07:615
PA	3:06:090	SCY	2:26:730	YYL	5:08:218	NT	3:22:826
AL	1:09:558	CYW	4:31:007	CY	1:49:531	NY	2:11:512
DF	3:11:382	HHY	4:19:300	LX	4:36:523	SA	2:58:048
MD	1:09:703	SMI	2:10:035	ZLE	1:47:298	UK	2:09:287
SU	1:55:266	WCF	2:40:746	ZLI	1:17:922	YE	3:59:356

Table B.1: Recording duration of the subjects.

APPENDIX C

DISTRIBUTION OF PROSODIC BOUNDARIES

Tables C.1 to C.4 show the distribution of prosodic breaks in the four languages.

			Position			
Hierarchy	P_{-2}	P_{-1}	P_0	P_1	P_2	Total
DSP0						
BI0	17	12	2	29	17	77
BI1	73	61	6	64	66	270
BI3	2	0	6	5	5	18
BI4	2	1	98	6	5	112
none	18	38	0	8	19	83
Subtotal	112	112	112	112	112	560
DSP1						
BI0	23	9	3	26	23	84
BI1	65	77	5	84	85	316
BI3	3	0	4	7	3	17
BI4	5	0	114	7	5	131
none	30	40	0	2	10	82
Subtotal	126	126	126	126	126	630
DSP2						
BI0	5	4	0	6	8	23
BI1	23	20	0	21	32	96
BI3	1	0	0	6	1	8
BI4	2	0	45	9	1	57
none	14	21	0	3	3	41
Subtotal	45	45	45	45	45	225
Total	283	283	283	283	283	1415

Table C.1: Distribution of prosodic boundaries in English. 'none' indicates there is no BI applicable to the syllable break due to an absence of a word boundary.

			Position			
Hierarchy	P_{-2}	P_{-1}	P_0	P_1	P_2	Total
DSP0						
BI0	15	20	0	39	21	95
BI1	194	207	14	176	159	750
BI2	14	6	98	9	30	157
BI3	8	0	61	8	15	92
BI4	1	0	52	0	7	60
BI5	1	0	8	1	1	11
Subtotal	233	233	233	233	233	1165
DSP1						
BI0	6	27	0	38	23	94
BI1	177	183	0	164	137	661
BI2	18	3	30	9	25	85
BI3	7	0	54	2	22	85
BI4	3	0	97	0	5	105
BI5	2	0	32	0	1	35
Subtotal	213	213	213	213	213	1065
DSP2						
BI0	3	5	0	12	5	25
BI1	46	49	0	38	33	166
BI2	4	1	3	2	9	19
BI3	2	1	5	3	4	15
BI4	0	0	35	1	4	40
BI5	1	0	13	0	1	15
Subtotal	56	56	56	56	56	280
Total	502	502	502	502	502	2510

Table C.2: Distribution of prosodic boundaries in Guoyu.

			Position			
Hierarchy	P_{-2}	P_{-1}	P_0	P_1	P_2	Total
DSP0						
BI0	12	8	0	42	19	81
BI1	128	204	3	149	86	570
BI2	94	30	100	45	119	388
BI3	7	0	111	6	13	137
BI4	1	0	26	0	4	31
BI5	0	0	2	0	1	3
Subtotal	242	242	242	242	242	1210
DSP1						
BI0	8	5	0	54	19	86
BI1	92	175	1	115	66	449
BI2	102	30	41	33	99	305
BI3	8	0	90	7	19	124
BI4	0	0	66	0	6	72
BI5	0	0	12	1	1	14
Subtotal	210	210	210	210	210	1050
DSP2						
BI0	5	1	0	13	2	21
BI1	27	47	0	39	23	136
BI2	23	8	3	2	23	59
BI3	1	0	15	2	6	24
BI4	0	0	26	0	1	27
BI5	0	0	12	0	1	13
Subtotal	56	56	56	56	56	280
Total	508	508	508	508	508	2540

Table C.3: Distribution of prosodic boundaries in Putonghua.

			Position			
Hierarchy	P_{-2}	P_{-1}	P_0	P_1	P_2	Total
DSP0						
BI0	0	0	0	0	0	0
BI1	22	28	2	7	45	104
BI2	3	0	1	0	0	4
BI3	2	0	90	6	8	106
BI4	0	0	2	0	0	2
none	68	67	0	82	42	259
Subtotal	95	95	95	95	95	475
DSP1						
BI0	0	0	0	0	0	0
BI1	19	12	0	9	30	70
BI2	1	0	0	0	2	3
BI3	2	0	72	10	10	94
BI4	0	0	33	0	0	33
none	83	93	0	86	63	325
Subtotal	105	105	105	105	105	525
DSP2						
BI0	0	0	0	0	0	0
BI1	7	3	0	1	13	24
BI2	0	0	0	1	1	2
BI3	1	0	18	7	2	28
BI4	0	0	34	0	0	34
none	44	49	0	43	36	172
Subtotal	52	52	52	52	52	260
Total	252	252	252	252	252	1260

Table C.4: Distribution of prosodic boundaries in Japanese. ‘none’ indicates there is no BI applicable to the syllable break due to an absence of a word boundary.

APPENDIX D

CONTRIBUTIONS OF EACH SUBJECT IN PROSODIC BREAK ANALYSES

Tables D.1 through D.4 show the number of discourse disjunctures each subject contributed in prosodic break analyses in this study. The languages concerned are English, Guoyu, Putonghua, and Japanese, accordingly.

ID	DSP0	DSP1	DSP2	Total
Male				
BH	20	18	10	48
JH	24	12	5	41
MK	16	21	2	39
PA	13	18	6	37
Subtotal	73	69	23	165
Female				
AL	3	7	4	14
DF	19	31	5	55
MD	7	8	6	21
SU	10	11	7	28
Subtotal	39	57	22	118
Total	112	126	45	283

Table D.1: Distribution of the number of cases of discourse disjunctures each subject contributed in prosodic break analyses in English.

ID	DSP0	DSP1	DSP2	Total
Male				
CHL	9	14	6	29
HPH	65	61	11	137
SCH	23	16	7	46
SCY	17	18	6	41
Subtotal	114	109	30	253
Female				
CYW	34	29	6	69
HHY	44	28	8	80
SMI	22	22	5	49
WCF	19	25	7	51
Subtotal	119	104	26	249
Total	233	213	56	502

Table D.2: Distribution of the number of cases of discourse disjunctures each subject contributed in prosodic break analyses in Guoyu.

ID	DSP0	DSP1	DSP2	Total
Male				
FBS	28	24	5	57
LJG	24	32	6	62
XD	30	30	7	67
YYL	54	39	12	105
Subtotal	136	125	30	291
Female				
CY	15	22	8	45
LX	73	42	7	122
ZLE	9	11	6	26
ZLI	9	10	5	24
Subtotal	106	85	26	217
Total	242	210	56	508

Table D.3: Distribution of the number of cases of discourse disjunctures each subject contributed in prosodic break analyses in Putonghua.

ID	DSP0	DSP1	DSP2	Total
Male				
FY	12	12	8	32
IA	6	8	5	19
MS	16	22	8	46
NT	16	14	7	37
Subtotal	50	56	28	134
Female				
NY	5	8	6	19
SA	8	13	5	26
UK	4	13	6	23
YE	28	15	7	50
Subtotal	45	49	24	118
Total	95	105	52	252

Table D.4: Distribution of the number of cases of discourse disjunctures each subject contributed in prosodic break analyses in Japanese.

APPENDIX E

CONTRIBUTIONS OF EACH SUBJECT IN PITCH PEAK TREND ANALYSES

Tables E.1 through E.4 show the number of discourse disjunctures each subject contributed in pitch peak trend analyses in this study. The languages concerned are English, Guoyu, Putonghua, and Japanese, accordingly.

ID	DSP0	DSP1	DSP2	Total
Male				
BH	17	25	10	52
JH	26	19	7	52
MK	16	28	5	49
PA	16	20	7	43
Subtotal	75	92	29	196
Female				
AL	4	10	7	21
DF	21	35	8	64
MD	6	9	7	22
SU	13	16	7	36
Subtotal	44	70	29	143
Total	119	162	58	339

Table E.1: Distribution of the number of cases of discourse disjunctures each subject contributed in pitch peak trend analyses in English.

ID	DSP0	DSP1	DSP2	Total
Male				
CHL	9	14	6	29
HPH	65	60	11	136
SCH	23	16	7	46
SCY	18	18	6	42
Subtotal	115	108	30	253
Female				
CYW	35	30	6	71
HHY	36	25	6	67
SMI	25	24	5	54
WCF	17	25	5	47
Subtotal	113	104	22	239
Total	228	212	52	492

Table E.2: Distribution of the number of cases of discourse disjunctures each subject contributed in pitch peak trend analyses in Guoyu.

ID	DSP0	DSP1	DSP2	Total
Male				
FBS	26	22	4	52
LJG	21	30	5	56
XD	30	28	7	65
YYL	51	37	11	99
Subtotal	128	117	27	272
Female				
CY	14	19	6	39
LX	72	41	6	119
ZLE	8	11	6	25
ZLI	9	10	5	24
Subtotal	103	81	23	207
Total	231	198	50	479

Table E.3: Distribution of the number of cases of discourse disjunctures each subject contributed in pitch peak trend analyses in Putonghua.

ID	DSP0	DSP1	DSP2	Total
Male				
FY	11	11	5	27
IA	4	7	4	15
MS	13	25	8	46
NT	18	14	7	39
Subtotal	46	57	24	127
Female				
NY	7	9	7	23
SA	7	14	4	25
UK	5	14	5	24
YE	28	14	7	49
Subtotal	47	51	23	121
Total	93	108	47	248

Table E.4: Distribution of the number of cases of discourse disjunctures each subject contributed in pitch peak trend analyses in Japanese.

APPENDIX F

CONTRIBUTIONS OF EACH SUBJECT IN DURATION ANALYSES

Tables F.1 through F.4 show the number of discourse disjunctures each subject contributed in duration analyses in this study. The languages concerned are English, Guoyu, Putonghua, and Japanese, accordingly.

ID	DSP0	DSP1	DSP2	Total
Male				
BH	4	13	0	17
JH	8	17	6	31
MK	9	17	5	31
PA	6	13	7	26
Subtotal	27	60	18	105
Female				
AL	1	7	7	15
DF	14	27	7	48
MD	3	10	7	20
SU	2	10	5	17
Subtotal	20	54	26	100
Total	47	114	44	205

Table F.1: Distribution of the number of cases of discourse disjunctures each subject contributed in duration analyses in English.

ID	DSP0	DSP1	DSP2	Total
Male				
CHL	8	14	5	27
HPH	36	48	11	95
SCH	2	7	4	13
SCY	10	15	6	31
Subtotal	56	84	26	166
Female				
CYW	17	24	6	47
HHY	19	21	8	48
SMI	3	13	4	20
WCF	8	19	7	34
Subtotal	47	77	25	149
Total	103	161	51	315

Table F.2: Distribution of the number of cases of discourse disjunctures each subject contributed in duration analyses in Guoyu.

ID	DSP0	DSP1	DSP2	Total
Male				
FBS	16	16	3	35
LJG	5	16	5	26
XD	4	12	4	20
YYL	34	31	12	77
Subtotal	59	75	24	158
Female				
CY	3	9	6	18
LX	47	31	3	81
ZLE	4	4	1	9
ZLI	2	4	2	8
Subtotal	56	48	12	116
Total	115	123	36	274

Table F.3: Distribution of the number of cases of discourse disjunctures each subject contributed in duration analyses in Putonghua.

ID	DSP0	DSP1	DSP2	Total
Male				
FY	6	6	2	14
IA	3	5	3	11
MS	8	19	6	33
NT	3	3	1	7
Subtotal	20	33	12	65
Female				
NY	4	8	7	19
SA	4	11	5	20
UK	1	9	5	15
YE	6	9	4	19
Subtotal	15	37	21	73
Total	35	70	33	138

Table F.4: Distribution of the number of cases of discourse disjunctures each subject contributed in duration analyses in Japanese.

LIST OF REFERENCES

Beckman, M. E. (1996). The parsing of prosody. *Language and Cognitive Processes, 11*(1/2), 17–67.

Beckman, M. E., & Ayers Elam, G. (1997). *Guidelines for ToBI labelling*. Unpublished manuscript, The Ohio State University.

Beckman, M. E., & Pierrehumbert, J. (1986). Intonation structure in English and Japanese. *Phonology Yearbook, 3*, 255–310.

Berkovits, R. (1984). A perceptual study of sentence-final intonation. *Language and Speech, 27*(4), 291–308.

Bestgen, Y. (1998). Segmentation markers as trace and signal of discourse structure. *Journal of Pragmatics, 29*, 753–763.

Campbell, W. N., & Isard, S. D. (1991). Segment durations in a syllable frame. *Journal of Phonetics, 19*(1), 37–47.

Carlson, K., Clifton, C., Jr., & Frazier, L. (2001). Prosodic boundaries in adjunct attachment. *Journal of Memory and Language, 45*(1), 58–81.

Chafe, W. L. (1980). *The pear stories: Cognitive, cultural, and linguistic aspects of narrative production. Advances in discourse processes, 3*. Norwood, NJ: Ablex.

Chao, Y.-R. (1968). *A grammar of spoken Chinese*. Berkeley & Los Angeles: University of California.

Clancy, P. M. (1980). Referential choice in English and Japanese narrative discourse. In W. L. Chafe (Ed.), *The pear stories: Cognitive, cultural, and linguistic aspects of narrative production. Advances in discourse processes, 3* (pp. 127–202). Norwood, NJ: Ablex.

Cooper, W. E., Paccia, J. M., & Lapointe, S. G. (1978). Hierarchical coding in speech timing. *Cognitive Psychology, 10*, 154–177.

Cooper, W. E. & Sorensen, J. M. (1977). Fundamental frequency contours at syntactic boundaries. *Journal of the Acoustical Society of America, 62*, 683–692.

Cutler, A., & Carter, D. M. (1987). The predominance of strong initial syllables in the English vocabulary. *Computer Speech and Language, 2*(1), 133–142.

Erbaugh, M. (1990). Mandarin oral narratives compared with English: the pear/guava stories. *Journal of Chinese Language Teacher's Association, 25* (2), 21–42.

Ferreira, F. (1993). Creation of prosody during sentence production. *Psychological Review, 100* (2), 233–253.

Fisher, C., & Tokura, H. (1996). Acoustic cues to grammatical structure in infant-directed speech: Cross-linguistic evidence. *Child Development, 67*(6), 3192–3218.

Flanigan, B. O., & Norris, F. P. (2000). Cross-dialectal comprehension as evidence for boundary mapping: Perceptions of the speech of southeastern Ohio. *Language Variation and Change, 12*, 175–201.

Fon, J., & Chiang, W.-Y. (1999). What does Chao have to say about tones? – A case study of Taiwan Mandarin. *Journal of Chinese Linguistics, 27*(1), 13–37.

Fon, J. & Johnson, K. (2000). Speech timing patterning as an indicator of discourse and syntactic boundaries. *Proceeding of 6th International Conference on Spoken Language Processing, 2*, 555–558.

Fougeron, C., & Keating, P. A. (1997). Articulatory strengthening at edges of prosodic domains. *Journal of the Acoustical Society of America, 101*, 3728–3740.

Fowler, C. A., Levy, E. T., & Brown, J. M., (1998). Reductions of spoken words in certain discourse contexts. *Journal of Memory and Language, 37*(1), 24–40.

Goldman-Eisler, F. (1972). Pauses, clauses, sentences. *Language and Speech, 15*, 103–113.

Grosz, B., & Hirschberg, J. (1992). Some intonational characteristics of discourse structure. *Proceedings of the 2nd International Conference on Spoken Language Processing*, 429–432.

Grosz, B., & Sidner, C. L. (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics*, *12*(3), 175–204.

Heeman, P., & Allen, J. F. (1995, June). The Trains spoken dialog corpus, CD-ROM, Linguistics Data Consortium.

Hirschberg, J., & Nakatani, C. H. (1996). A prosodic analysis of discourse segments in direction-giving monologues. *Proceeding of the 4th International Congress of Spoken Language Processing*, 286–293.

Huggins, A. W. F. (1974). An effect of syntax on syllable timing. *Quarterly Progress Report, MIT*, *114*, 179–185.

Jusczyk, P. W., Houston, D. M., & Newsome, M. (1999). The beginnings of word segmentation in English-learning infants. *Cognitive Psychology*, *39*, 159–207.

Kärkkäinen, E. (1996). Preferred argument structure and subject role in American English conversational discourser. *Journal of Pragmatics*, *25* (5), 675–701.

Katz, W. F., Beach, C. M., Jenouri, K., & Verma, S. (1996). Duration and fundamental frequency correlates of phrase boundaries in productions by children and adults. *The Journal of the Acoustical Society of America*, *99*(5), 3179–3191.

Kemler Nelson, D. G., Hirsh-Pasek, K., Jusczyk, P. W., & Cassidy, K. W. (1989). How the prosodic cues in motherese might assist language learning. *Journal of Child Language*, *16*(1), 55–68.

Klatt, D. H. (1975). Vowel lengthening is syntactically determined in a connected discourse. *Journal of Phonetics*, *3*, 129–140.

Koopmans-van Beinum, F. J., & van Donzel, M. E. (1996). Relationship between discourse structure and dynamic speech rate. *Proceeding of the 4th International Congress of Spoken Language Processing*, 1724–1727.

Kreiman, J. (1982), Perception of sentence and paragraph boundaries in natural conversation. *Journal of Phonetics*, *10*(2), 163–175.

Ladd, D. R. (1988). Declination 'reset' and the hierarchical organization of utterances. *The Journal of the Acoustical Society of America*, *84*(2), 530–544.

Lehiste, I. (1971). Temporal organization of spoken language. In L. L. Hammerich, R. Jakobson, & E. Zwirner (Eds.), *Form and substance: Phonetic and linguistic papers presented to Eli Fischer-Jørgensen* (pp. 159–169). Copenhagen, Denmark: Akademisk Forlag.

Lehiste, I. (1973). Phonetic disambiguation of syntactic ambiguity. *Glossa*, *7*(2), 107–121.

Lehiste, I. (1975a). The phonetic structure of paragraphs. In A. Cohen & S. Nooteboom (Eds.), *Structure and process in speech perception* (pp. 195–206). Heidelberg, Germany: Springer-Verlag.

Lehiste, I. (1975b). The role of temporal factors in the establishment of linguistic units and boundaries. In W. U. Dressler, & F. V. Mares (Eds.), *Phonologica 1972* (pp. 115–122). München-Salzburg, Germany: Wilhelm Fink Verlag.

Lehiste, I. (1979a). Perception of sentence and paragraph boundaries. In B. Lindblom & S. Ohman (Eds.), *Frontiers of speech communication research* (pp. 191–201). New York: Academic Press.

Lehiste, I. (1979b). Sentence boundaries and paragraph boundaries—perceptual evidence. In P. R. Clyne, W. F. Hanks, & C. L. Hofbauer (Eds.) *The elements: A parasession on linguistic units and levels* (pp. 99–201). Chicago: The Chicago Linguistic Society.

Lehiste, I., Olive, J. P, & Streeter, L. A. (1976). Role of duration in disambiguating syntactically ambiguous sentences. *Journal of the Acoustical Society of America*, *60*(5), 1199–1202.

Lehiste, I., & Wang, W. S.-Y. (1977). Perception of sentence and paragraph boundaries with and without semantic information. In W. U. Dressler & O. E. Pfeiffer (Eds.), *Phonologica 1976* (pp. 277–283). Innsbruck: Institut fur Sprachwissenschaft der Universität Innsbruck.

Li, C. N., & Thompson, S. A. (1981). *Mandarin Chinese: A functional reference grammar*. Berkeley & California, LA: University of California Press.

Maclay, H., & Osgood, C. E. (1959). Hesitation phenomena in spontaneous English speech. *Word*, *15*, 19–44.

Marr, D. (1982). *Vision*. San Francisco: Freeman.

Merriam-Webster online collegiate dictionary. (2002). Retrieved April 3, 2002, from http://www.webster.com

Nakatani, L. H., O'Connor, K. D., & Aston, C. H. (1981). Prosodic aspects of American English speech rhythm. *Phonetica*, *38*, 84–106.

Nazzi, T., Bertoncini, J., & Mehler, J. (1998). Language discrimination by newborns: Towards an understanding of the role of rhythm. *Journal of Experimental Psychology: Human Perception and Performance*, *24*, 756–766.

Oller, D. K. (1973). The duration of speech segments: The effect of position in utterance and word length. *Journal of the Acoustical Society of America*, *54*(5), 1235–1247.

O'Malley, M. H., Kloker, O. R., & Dara-Abrams, D. (1973). Recovering parentheses from spoken algebraic expressions. *I.E.E.E. Transactions on Audio and Electroacoustics*, *AU–21*(3), 217–220.

Pakosz, M., & Flaschner, V. (1988). Prosodic features and narrative strategies in Polish discourse. *Papers and Studies in Contrastive Linguistics*, *24*, 33–46.

Peng, S.-H., Chan, M. K. M., Tseng, C.-Y., Huang, T., Lee, O. J., & Beckman, M. (2000). *A pan-Mandarin ToBI*. Unpublished manuscript, The Ohio State University.

Price, P. J., Ostendorf, M., Shattuck-Hufnagel, S., & Fong, C. (1991). The use of prosody in syntactic disambiguation. *The Journal of the Acoustical Society of America*, *90*(6), 2956–2970.

Quené, H. (1992). Durational cues for word segmentation in Dutch. *Journal of Phonetics*, *20*(3), 331–350.

Quené, H. (1993). Segment duration and accent as cues to word segmentation in Dutch. *The Journal of the Acoustical Society of America*, *94*(4), 2027–2035.

Ratner, N. B. (1986). Durational cues which mark clause boundaries in mother-child speech. *Journal of Phonetics*, *14*(2), 303–309.

Raymond, W. D., Makashay, M. J., Dautricourt, R., Johnson, K., Hume, E., & Pitt, M. (2001, December). Variation in conversation: An introduction to the Buckeye Speech Corpus. Poster session presented at the annual meeting of the Acoustical Society of America, Ft. Lauderdale, FL.

Rotondo, J. A. (1984). Clustering analyses of subjective partitions of text. *Discourse Processes*, *7*, 69–88.

Sanderman, A. A., & Collier, R. (1996). Prosodic rules for the implementation of phrase boundaries in synthetic speech. *The Journal of the Acoustical Society of America*, *100*(5), 3390–3397.

Schaffer, D. (1984). The role of intonation as a cue to topic management in conversation. *Journal of Phonetics*, *12*(4), 327–344.

Scott, D. R. (1982). Duration as a cue to the perception of a phrase boundary. *Journal of the Acoustical Society of America*, *71*(4), 996–1007.

Selkirk, E. O. (1984). *Phonology and syntax: The relation between sound and structure*. Cambridge, MA: MIT Press.

Selkirk, E. O. (1986). On derived domains in sentence phonology. *Phonology Yearbook*, *3*, 371–405.

Shen, X. S. (1992). A pilot study on the relation between the temporal and syntactic structures in Mandarin. *Journal of the International Phonetic Association*, *22*(1-2), 35–43.

Shriberg, E. E. (1994). *Preliminaries to a theory of speech disfluencies*. Unpublished doctoral dissertation, University of California at Berkeley.

Snow, D. (1994). Phrase-final syllable lengthening and intonation in early child speech. *Journal of Speech and Hearing Research*, *37*(4), 831–840.

Speer, S. R., Kjelgaard, M. M., & Dobroth, K. M. (1996). The influence of prosodic structure on the resolution of temporary syntactic closure ambiguities. *Journal of Psycholinguistic Research*, *25*(2), 249–271.

Streeter, L. A. (1978). Acoustic determinants of phrase boundary perception. *Journal of the Acoustical Society of America*, *64*(6), 1582–1592.

Swerts, M. (1997). Prosodic features at discourse boundaries of different strength. *The Journal of the Acoustical Society of America*, *101*(1), 514–521.

Swerts, M. (1998). Filled pauses as markers of discourse structure. *Journal of Pragmatics*, *30*(4), 485–496.

Swerts, M., & Geluykens, R. (1994). Prosody as a marker of information flow in spoken discourse. *Language and Speech*, *37*(1), 21–43.

Terken, J.M.B. (1984). The distribution of pitch accents as a function of discourse structure. *Language and Speech*, *27*(3), 269–289.

Traum, D. R., & Heeman, P. A. (1997). Utterance units in spoken dialogue. *Dialogue Processing in Spoken Language Systems*, *1236*, 125–140.

Tseng, C.-Y., & Chou, F.-C. (1999). Machine readable phonetic transcription system for Chinese dialects spoken in Taiwan. *The Journal of the Acoustical Society of Japan (E)*, *20* (3), 215–223.

Venditti, J. J. (1997). Japanese ToBI labelling guidelines. In K. Ainsworth-Darnell & M. D'Imperio (Eds.), *Papers from the Linguistics Laboratory. The Ohio State University Working Papers in Linguistics*, *50*, 127–162.