TYPOLOGIZING NATIVE LANGUAGE INFLUENCE ON INTONATION IN A SECOND LANGUAGE:
THREE TRANSFER PHENOMENA IN JAPANESE EFL LEARNERS

Aaron Lee Albin

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements for the degree
Doctor of Philosophy
in the Department of Linguistics
and the Department of Second Language Studies,
Indiana University
July 2015
Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Doctoral Committee

__________________________________________
Kenneth de Jong, Ph.D.

__________________________________________
Isabelle Darcy, Ph.D.

__________________________________________
Yoshihisa Kitagawa, Ph.D.

__________________________________________
Markus Dickinson, Ph.D.

May 14, 2015
For Nina
Acknowledgments

I would like to thank the following individuals and organizations for their support leading up to and throughout my graduate career.

Committee (For being the best mentors anyone could ask for): Ken de Jong, Isabelle Darcy, Yoshihisa Kitagawa, Markus Dickinson.

Departments (For supporting me through seven years of graduate school): Linguistics, Second Language Studies.

Faculty (For giving me a world-class graduate education): Damir Cavar, Erik Willis, Jeff Holliday, John Kruschke, Kathleeen Bardovi-Harlig, Keiko Kuriyama, Kelley Berkson, Lisa Gershkoff, Malgorzata Cavar, Misako Matsubara, Robert Botne, Stuart Davis, Tessa Bent.

Researchers (For lively discussions and sharing your research): Andrew Rosenberg, Becky Taylor, Chiharu Tsurutani, Ineke Mennen, Janet Pierrehumbert, Jennifer Venditti, Julia McGory, Kikuo Maekawa, Laura Dilley, Luca Rognoni, Maria Busà, Mary Beckman, Monika Zięba-Plebankiewicz, Nadia Bouchhioua, Sabine Zerbian, Seth Goss, Shin'ichirō Ishihara, Sun-Ah Jun, Takehiko Makino, Ulrike Gut, and many others.

Audiences (For helping shape the ideas that went into this dissertation): Experimental and Theoretical Advances in Prosody 3, Linguistics Department Colloquium, Second Language Psycholinguistic Lab, Second Language Studies Colloquium.
Technical (For providing the tools that made this dissertation possible): Indiana University IT Training, Indiana University Libraries, Jefferson Davis, National Institute of Informatics Speech Resources Consortium, Nobuaki Minematsu, Paul Boersma, R Core Team, Santiago Barreda.

Financial (For supporting me semester after semester): Charles Lin, College of Arts and Sciences, Department of Linguistics, East Asian Languages and Cultures, East Asian Studies Center, Keiko Kuriyama, Ken de Jong, Pan-Asia Institute, Yoshihisa Kitagawa.


Omaha (For getting me where I am today): Anne Ludwig, Debbie Albin, Frank Bramlett, James Albin, Jamie Albin, Jimmy Albin, Lori Arias, Lorraine Novotny, Owen Mordaunt, Russ Albin, Stan Novotny, University of Nebraska at Omaha.

And finally, most of all, to Yoshino - for your undying patience through it all.

This dissertation was supported under a Dissertation Completion fellowship from the College of Arts and Sciences at Indiana University - Bloomington.
While a substantial body of research has accumulated regarding how intonation is acquired in a second language (L2), the topic has historically received relatively little attention from mainstream models of L2 phonology. As such, a unified theoretical framework suited to address unique acquisitional challenges specific to this domain of L2 knowledge (such as form-function mapping) has been lacking. The theoretical component of the dissertation makes progress on this front by taking up the issue of crosslinguistic transfer in L2 intonation. Using Mennen's (2015) L2 Intonation Learning theory as a point of departure, the available empirical studies are synthesized into a typology of the different possible ways two languages' intonation systems can mismatch as well as the concomitant implications for transfer.

Next, the methodological component of the dissertation presents a framework for overcoming challenges in the analysis of L2 learners' intonation production due to the interlanguage mixing of their native and L2 systems. The proposed method involves first creating a stylization of the learner's intonation contour and then running queries to extract phonologically-relevant features of interest for a particular research question. A novel approach to stylization is also introduced that not only allows for transitions between adjacent pitch targets to have a nonlinear shape but also explicitly parametrizes and stores this nonlinearity for analysis.
Finally, these two strands are integrated in a third, empirical component to the dissertation. Three kinds of intonation transfer, representing nodes from different branches of the typology, are examined in Japanese learners of English as a Foreign Language (EFL). For each kind of transfer, fourteen sentences were selected from a large L2 speech corpus (English Speech Database Read by Japanese Students), and productions of each sentence by approximately 20-30 learners were analyzed using the proposed method. Results suggest that the three examined kinds of transfer are stratified into a hierarchy of relative frequency, with some phenomena occurring much more pervasively than others.

Together as a whole, the present dissertation lays the groundwork for future research on L2 intonation by not only generating empirical predictions to be tested but also providing the analytical tools for doing so.

__________________________________________
Kenneth de Jong, Ph.D.

__________________________________________
Isabelle Darcy, Ph.D.

__________________________________________
Yoshihisa Kitagawa, Ph.D.

__________________________________________
Markus Dickinson, Ph.D.
# TABLE OF CONTENTS

**Chapter 1: Introduction** ................................................................. 1

1.1 **Background** ................................................................. 1

1.1.1 Transfer in second language acquisition 2
1.1.2 Intonational phonology 5
1.1.3 The acquisition of intonation in a second language 10
1.1.4 The role of transfer in L2 intonation 13

1.2 **Three components of the dissertation** 15

1.2.1 Theoretical component 16
1.2.2 Methodological component 17
1.2.3 Empirical component 20

1.3 **Preview of the rest of the dissertation** 24

**Chapter 2: Typology** ................................................................. 26

2.1 **Position** ................................................................. 31

2.1.1 Emptiness 32
2.1.2 Fixedness 33

2.2 **Category** ................................................................. 35

2.2.1 Presence 35
2.2.2 Conditioning 36

2.3 **Accentuation** ............................................................... 37

2.3.1 Flexibility 38
2.3.2 Word types 39

2.4 **Realization** ................................................................. 40

2.4.1 Alignment 41
2.4.2 Scaling 42

2.5 **Function** ................................................................. 44

2.5.1 Expression 45
2.5.2 Specific form 46

2.6 **Density** ................................................................. 48

2.6.1 Pitch accents 48
2.6.2 Edge tones 50
Chapter 3: Methodological framework ................................................................. 52

3.1 Adopted methods for F0 extraction and visualization 52

3.1.1 Choice of F0 tracker 53
3.1.2 Parameter settings 56
3.1.3 Visualization of F0 data 58

3.2 Stylization 66

3.3 Vertices 69

3.3.1 What vertices represent 69
3.3.2 Determining vertex locations 70

3.4 Transitions 72

3.4.1 Nonlinear vs. linear stylization 72
3.4.2 Distribution chosen for describing transition shape 75
3.4.3 Re-parametrization in terms of threshold and gradience 79
3.4.4 Automatically fitting transition curves to actual F0 data 83

3.5 Storing and analyzing stylization data 89

Chapter 4: Research questions ................................................................. 93

4.1 Description of the intonation system of Japanese and English 93

4.1.1 Broad overview of the AM model of Japanese intonation 94
4.1.2 Broad overview of the AM model of English intonation 101

4.2 Research questions 107

4.2.1 Research question 1 108
4.2.2 Research question 2 112
4.2.3 Research question 3 115

Chapter 5: Procedure ................................................................................. 123

5.1 Description of the dataset 123

5.1.1 Overview of corpus as a whole 123
5.1.2 Corpus structure 126
5.1.3 Elicitation methods and rationale 128
5.1.4 Selected sub-corpus 131

5.2 Selection of sentences 135

5.2.1 Sentences selected for Research Question 1 137
Chapter 1: Introduction

The present dissertation accomplishes three different research goals. Chapter 2 proposes a typology of the ways in which two languages' intonational phonologies can differ (and therefore lead to negative transfer in second language acquisition). Chapter 3 establishes a novel methodological framework for coding and analyzing intonation contours in a way that separates surface description from theoretical interpretation - a distinction that is especially crucial when dealing with second language learners. The remainder of the dissertation (Chapters 4-7) analyses production data from a large speech corpus in order to determine which of three kinds of crosslinguistic intonation transfer occurs the most frequently.

This chapter sets the stage for the dissertation in three ways. First, section 1.1 presents various pieces of background information relevant to the dissertation concerning intonation and second language acquisition. Next, section 1.2 lays out the three goals (theoretical, empirical, and methodological) that the dissertation seeks to accomplish. Finally, section 1.3 previews the structure of the rest of the dissertation.

1.1 Background

This section reviews four strands of previous literature, namely: cross-linguistic transfer in second language acquisition (§1.1.1), intonational phonology (§1.1.2), the acquisition of intonation in a second language (§1.1.3), and the role of transfer in second language intonation
The intent of this section is not to give a comprehensive overview of the vast body of research that has been conducted on these topics. Rather, only the core information about each topic will be presented, with the goal of clarifying the assumptions about these topics that will underlie the rest of the dissertation.

1.1.1 Transfer in second language acquisition

Transfer (also referred to as (language/linguistic/crosslinguistic) 'interference' or 'influence') refers to the process whereby a multilingual individual consciously or unconsciously applies knowledge from one of their languages during the process of learning or using (i.e. reading, writing, speaking, or listening) another one of their languages (Weinreich 1953). This kind of language mixing can take many forms. In a bilingual, for instance, it may involve cases where one's experience in having learned a second language (L2) affects one's use of a first language (L1). Conversely, the crosslinguistic influence can occur in the other direction, with one's L1 affecting (or 'slipping into') performance in an L2. Speakers of three or more languages exhibit many more possible directions for transfer to occur.

Transfer can sometimes be beneficial ('positive transfer'). For example, if two languages are identical in some regard, then the ability to borrow one's pre-existing knowledge of one language can helpfully speed along the acquisition process of the other. In other situations, wholesale application of knowledge from the other language can be detrimental ('negative transfer'). For example, if two languages are different in some regard, then transfer from one language would likely create an illicit structure in the other. This may be especially common if two languages are superficially similar and only different in the finer details. See Ellis

Of this broader space of possible kinds of transfer, the present dissertation deals only with negative transfer from an L1 into an L2, with special attention to the case of classroom-instructed sequential bilinguals. For terminological simplicity, the shorthand term 'transfer' will be used throughout the rest of the dissertation. This should be interpreted as referring more specifically to 'L1-to-L2 negative transfer'.

Prototypical cases of transfer occur where a learner's limited capabilities in the L2 prevents them from successfully performing some kind of task in the L2, hence their only recourse is to 'borrow' knowledge from the L1 in order to fill that gap in their knowledge. Under this account, since the root cause is the learner's limited capabilities in the L2, transfer is most frequent in the earliest stages of L2 acquisition, where L2 knowledge is most lacking. Later, as the learner gains proficiency in the L2, transfer becomes less and less necessary and therefore decreases in frequency. However, the fact that even highly advanced L2 learners fail to become fully targetlike suggests that transfer often does not disappear entirely (Moyer 1999).

From the perspective of research in second language acquisition, it is important to note that transfer is ultimately a researcher's interpretation of the data. As such, transfer itself cannot be measured, just the theoretical predictions thereof. Empirical evidence for transfer can take two

---

1 While prototypical cases of transfer involve some aspect of the L1 showing up in L2 production, it is also possible for the L1 to influence L2 production in another, more indirect way. Namely, some characteristic of the L1 can impact the range and distribution of L2 structures the learner uses. In the present dissertation, the term transfer is used in this broader sense as a cover term for both of these scenarios - that is, any kind of influence attributable to the L1.
forms. First, a learner could be observed to do something L1-like in the L2. Second, a learner could fail to do something L2-like in the L2 due to some difference between the L1 and the L2. In either case, since causality is being attributed to the L1, a systematic comparison of multiple L1 groups is necessary in order to obtain the strongest possible evidence for transfer.

The potentially idiosyncratic way in which a learner's combines L1 and L2 knowledge can be thought of as contributing to their interlanguage system at any given stage of development (Selinker 1972). However, transfer alone cannot account for every aspect of a learner's interlanguage. Learners can be non-targetlike in myriad ways, many of which are not predicted by transfer. Thus, an L2 utterance can be free of influences attributable to transfer from the L1 and still be non-targetlike in other ways. For example, a learner could differ from a native speaker in the fine phonetic details of producing an L2 sound, or produce L2 utterances at a slowed speech rate. Both of these are markers of one's status as an L2 learner but have nothing to do with transfer.

Moreover, even in cases where transfer is a valid explanation for something, there are often other equally valid alternative explanations that do not involve transfer, such as general universal principles or developmental factors similar to those occurring in L1 acquisition (Major 2008:75). For example, if something that appears at first to be transfer has also been attested in various other typologically diverse L1-L2 pairings for which transfer is not a valid explanation, then the relevant phenomenon may not be transfer after all. In many cases, transfer explanations and non-transfer explanations overlap - i.e., in some cases, both are plausible (and indeed both may be simultaneously operative). Thus, while transfer is an important and ubiquitous process in L2 acquisition, it must always be situated in a broader web of other inter-related factors. For
further discussion on what constitutes sufficient evidence for transfer and how it can be
disentangled from other competing explanations, see Jarvis (2000).

1.1.2 Intonational phonology

The present dissertation is concerned with the use of fundamental frequency (F0) in
speech. In the broadest possible sense, irrespective of any specific linguistic function, this may
be referred to as 'melody'. Note that this definition of melody excludes other aspects of speech
prosody such as temporal dynamics ('rhythm') or pausing. Each such dimension of the signal is
of course related to F0 but is independent enough to warrant an investigation in its own right.
The all-encompassing nature of this definition rightfully implies how multifaceted a construct
melody is. Indeed, melody may actually be best thought of as an epiphenomenon - a mere cover
term for various separate (but inter-related) smaller phenomena.

Of this larger picture, the present dissertation focuses more specifically on the use of F0
at the level above the word (i.e. for purposes other than lexical identity marking). While melody
is influenced (and, indeed, built on top of the building blocks of) lexically-contrastive tone,
stress, and accent, these are not the focus of the present dissertation. Instead, emphasis will be
placed on 'intonation' narrowly defined, i.e. the phonological category structure at the utterance
level to signal non-lexical information such as illocutionary force or information structure.

Of the dozens of existing approaches to intonation analysis, the present dissertation
adopts the Autosegmental-Metrical (AM) framework (also referred to as a 'theory', 'model', or
'approach'). The AM framework has been successfully applied to describe the intonational
phonology of a wide variety of typologically diverse languages (cf. Jun 2005 and Jun 2014) At
its core, the AM framework is a means to describe intonation in terms of phonological
representations. This is directly in line with much of the Second Language Phonology literature, which also deals with the issue of how phonological representations from two languages map onto each other (often despite fine-grained phonetic differences). Thus, since the task of the present dissertation is to examine how L2 learners acquire intonation in a second language, the representational nature of the AM framework makes it well-suited to the task. The following is a brief overview of what the AM framework entails; see Ladd (2009) and Gooden et al. (2009) for a full discussion. Analyses of the two languages examined in the present dissertation - English and Japanese - in the AM framework are provided in Chapter 4.

One core assumption of the AM framework is the separation of form and function, thus distinguishing it from other frameworks like the Parallel Encoding and Target Approximation (PENTA) framework (Xu 2005). Thus, it is possible to describe an utterance as consisting of a string of phonological categories, any of which may be used for one function (e.g. marking a question) in one language but a different function (e.g. marking contrastive focus) in another. Another important assumption of the AM framework is that the F0 contour contains an inter-mixing of both linguistic and paralinguistic information. The overall level ('high pitch' vs. 'low pitch') and span ('pitch range' being wide or narrow) of the F0 contour may be considered examples of the latter. In the AM framework, these two dimensions are seen as a 'backdrop' for the string of phonological categories that make up the contour, as they can be independently modulated while keeping the category string fixed.

The AM framework takes the F0 target, as manifested by a turning point in the F0 track, as the phonetic primitive, not unlike the Articulatory Phonology approach to segmental phonetics (Browman & Goldstein 1986). A dynamic movement (i.e. a rise or a fall) is seen as merely a
transition between two adjacent F0 targets (Pierrehumbert & Beckman 1988). Any given F0 target may vary gradationally in the finer details of its phonetic realization - most notably its alignment (i.e. temporal location) or scaling (i.e. height). However, the anchoring (or 'association') of an F0 target to a given prosodic unit (e.g. a syllable at some specific point in the utterance) is seen as phonological in nature.² Accordingly, discrete boundaries have been found between contrasting intonational categories in a wide range of perceptual tasks, including categorical perception tasks using synthetic continua (Pierrehumbert & Steele 1989, Dilley 2005, Vanrell et al. 2013). Thus, the intonational categories (and the category boundaries between them) are assumed to be a part of the native speaker's linguistic knowledge. The inventory of these categories varies from one language to the next; hence, any given language can be described as having an 'intonational phonology'.

Broadly speaking, the AM framework defines two different kinds of phonological category: edge tones and pitch accents. An edge tone, a term borrowed from Ladd (2009), is phonologically associated with the edge of a Prosodic constituent of some size, thereby marking its edge (hence the name). When attached to a lower-level prosodic constituent, e.g. a 'minor phrase' for a prosodic word, the edge tone is referred to as a phrase tone (or 'phrase accent') and its transcription label uses a dash (e.g. L-). If, instead, it is attached to a higher-level prosodic...

² Specific scaling values can also be phonological in nature. While earliest work in the AM framework assumed tones could be either phonologically high or low (H or L), potentially with contextual upstepping/downstepping in specific phonetic realizations thereof (e.g. !H or ¡L), it is now recognized that "tone specifications are inherently arranged on a scale from low (L) to high (H)" (Gooden et al. 2009:206). Such added flexibility is necessary to account for surface facts such as "the L tone of tone 3 ('dipping tone') being phonetically lower than the L of tone 2 ('rising tone')" in Mandarin Chinese (Gooden et al. 2009:411).
constituent, e.g. a 'major'/intonational phrase or the level of the entire utterance, the edge tone is referred to as a boundary tone and its transcription label uses a percent sign (e.g. L%).

Inside the prosodic constituents thus delimited, most languages also have additional tonal material at structurally prominent syllables, such as at a primary-stressed syllable in English. A tone associated to such locations is referred to as a pitch accent, and its transcription label uses an asterisk (e.g. L*). The last pitch accent in a phrase is referred to as the 'nuclear' pitch accent (or 'nucleus'). While this is defined purely in terms of prosodic phonology, this pitch accent plays a special role in the information structure of many languages. In cases where multiple F0 targets are associated with the same syllable, they are joined by a plus sign and the one most closely aligned to the syllable in question receives the star in the transcription.3 For example, a rise that begins inside at a stressed syllable and terminates (i.e. plateaus) in the following syllable may be transcribed L*+H. Such cases are said to involve a 'bitonal pitch accent', such that the two F0 targets (L* and H) cohere together and function collectively as one unit.

As pointed out by Ladd (2009), it is important to distinguish between the basic tenants of AM framework as a whole (such as those just outlined) and its instantiation in the model of any one particular language. This distinction makes it possible to acknowledge that the AM framework as a whole can be on the right track even if some specific analyses, such as that of English outlined in Pierrehumbert (1980), may be incorrect and require revision in some of the details. (Note, for example, the change in pitch accent inventory between Pierrehumbert (1980)

---

3 This dichotomous notational convention is now recognized as overly simplistic since it glosses over what is actually a much more complicated reality. See Arvaniti et al. (2000) for relevant discussion.
and Beckman & Ayers-Elam (1997), plus even further suggested revisions in Dilley & Heffner (2013), *inter alia.*

It is also important to distinguish the AM framework (as a theoretical paradigm) from any specific methodological implementations thereof. The most frequent way the AM framework is applied in practice is by using the family of Tones and Break Indices (ToBI) systems (Beckman et al. 2005) to holistically code production data as a string of transcription symbols. However, this is by no means the only alternative. Notable examples of alternative transcriptions systems that are outside of ToBI proper but nonetheless follow the AM framework include the Intonational Variation in English (IViE) system (Grabe 2001, Grabe 2004), the Transcription of Dutch Intonation (ToDI) (Gussenhoven 2005), and Rhythm and Pitch (RaP) (Breen et al. 2012). Even beyond transcription systems, the AM framework has also informed the design of countless psycholinguistic experiments. Here again, this distinction makes it possible to acknowledge that the AM framework as a whole can be on the right track even if ToBI transcription systems suffer from several inherent design flaws, as pointed out by numerous researchers (e.g., Wightman 2002, Martin 2012).

To sum up, then, the AM framework is best thought of as not an analysis of any one language, nor any particular method, but rather a set of underlying theoretical principles for guiding thinking about how intonation is phonologically structured. The details of the inventory for any given particular language are an empirical question, and methods for applying the AM framework in practice can always be subject to refinement and improvement.
1.1.3 The acquisition of intonation in a second language

As one component of phonological knowledge, intonation categories need to be acquired in much the same way that segmental categories and lexical prosody do. Thus, for example, even with a mastery of English vowels, consonants, and stress locations in individual words, an L2 learner can still be quite non-targetlike if they do not know where F0 targets should be placed over the course of an utterance. Moreover, struggling with intonation in a second language appears to be quite common for numerous reasons. Not the least of these is the fact that learners not only need to learn the phonological categories themselves but also how they are used in interaction (based on the complex mappings between intonational form and pragmatic function). Other reasons L2 intonation is so challenging include (1) the highly contextual nature of the form-function mapping, (2) imprecise metalinguistic awareness by the average native speaker, (3) inadequate treatment in language classrooms and textbooks, (4) a lack of representation in orthography, and (5) the fact a given L2 contour can often be 'parsed' in numerous ways (some of which may be fully in accordance with the L1 system).

Given that the acquisition of non-lexical aspects of intonation involves not only form but also function (thus distinguishing it from most segmental phonology), it is of theoretical interest for the field of second language phonology. Nonetheless, intonation has historically played only a minor role in the field as a whole, especially relative to the much larger literature on the L2 acquisition of prosodic lexical contrasts. (See, for example, Dupoux et al. (2008) and the references cited therein.) This fact is evidenced by the fact none of the major models of L2 phonology, e.g. the Native Language Magnet (NLM) model (Kuhl & Iverson 1995), the Speech Learning Model (SLM) (Flege 1995), the Perceptual Assimilation Model (PAM) L2 (Best &
Tyler 2007), or Direct Mapping of Acoustics to Phonology (DMAP) (Darcy et al. 2012) make any direct reference to intonation.4

This, of course, does not mean that these models are inapplicable to intonation. Quite the contrary, many researchers have made reference to these models in framing their studies and interpreting their results. Mennen (1999:556) for instance, argues that the SLM would treat an L2 Greek L* H L% yes/no question as 'similar' to an L1 Dutch declarative in cases where a rise-fall contour would be created on the utterance-final word in both languages. In contrast, an L2 Greek L* H L% contour with the nuclear pitch accent on the utterance-final word be 'new' in cases where it creates rise-fall on an unstressed syllable - something supposedly impossible in L1 Dutch. Similarly, Gili Fivela (2012) applies PAM L2 and finds that, due to phonetic similarities between the two, Italian EFL learners assimilated L2 English non-focal pitch accents to their L1 Italian correction-focus pitch accents (as evidenced by diminished performance in a discrimination task). Thus, as evidenced by these two example studies, the major models of L2 phonology can in principle be extended to apply to intonation.

However, despite the fact that research on L2 intonation is less well-known from the perspective of the field of L2 phonology as a whole, there is in fact a substantial body of literature on the topic, with many hundreds of published articles relating to the topic. The maturity of the field can be seen from fact that there exist entire monographs synthesizing

---

4 More recently, an offshoot of PAM (the Perceptual Assimilation Model for Suprasegmentals, or "PAM-S"), has been developed. For details, see So & Best (2014) and the references cited therein. While the model purports to describe perception in terms of one's native sentence intonation categories (or "i-Categories"), to date, the model has been restricted to lexical prosody - more specifically, the perception of Mandarin lexical tone in naïve Cantonese, Japanese, English, and French listeners. Thus, while PAM-S is an important first step, much yet remains to be worked out before such a model can be account for the perception of non-lexical intonation categories in typologically diverse languages.

From this larger literature, at least five models of L2 intonation acquisition have emerged, each attempting to describe a different aspect of the problem. One early study, Cruz-Ferreira (1987:118), describes the range of strategies learners apply when interpreting an intonation contour in terms of a flowchart, distinguishing 'familiar' from 'unfamiliar' structures and patterns. Zięba-Plebankiewicz (2007:117) describes how typological preferences and universal processes affect the 'intentions' and realizations of L2 tones, and how those cascade into perception and production. The Prosodic Focus Marking Acquisition Model (Baker 2010:226) consists of five hypotheses ("acoustic transfer", "acoustic L2 challenge", "any prominence location helps transfer", "relationship L2 challenge", and "frequency L2 challenge") predicting a hierarchy of accuracy at both the acoustic and phonological levels.

The fifth model - the L2 Intonation Learning theory ("LILt") - is of the greatest direct relevance to the present study. LILt was first presented in Mennen (2015) but based on previous work such as Mennen (2008). LILt aims to account for the difficulties that L2 learners encounter in producing L2 intonation. The cornerstone of the model is the distinction between four different 'dimensions' for characterizing similarities and differences between the intonation system of an L1 and an L2 and, therefore, help predict where non-targetlikeness is likely to occur.
<table>
<thead>
<tr>
<th>Dimension</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Systemic</td>
<td>the inventory and distribution of categorical phonological elements</td>
</tr>
<tr>
<td>Realizational</td>
<td>the phonetic implementation of these categorical elements</td>
</tr>
<tr>
<td>Semantic</td>
<td>the functionality of the categorical elements or tunes</td>
</tr>
<tr>
<td>Frequency</td>
<td>the frequency of use of the categorical elements</td>
</tr>
</tbody>
</table>

Table 1.1: Four dimensions of Mennen's (2015) L2 Intonation Learning theory (LILt)

Of the four dimensions, it is interesting to note that SLM and PAM-L2 / PAM-S only focus on the first two of these four dimensions - the phonological categories themselves and their phonetic realizations. As noted at the beginning of this section, the functional component of intonation (LILt's third dimension) is a unique attribute that makes it of theoretical interest for L2 phonology. The fourth dimension can be thought of as a higher-level bird's eye view of the distribution in aggregate output of the phonology. Its inclusion in LILt is based on studies such as Grabe (2004), who found that even if two dialects have the same inventory of categories, they may nonetheless differ in the frequency they are used. Taken together, these four dimensions make it clear that speaking entirely in terms of phonological categories and phonetic realization is too restrictive for L2 intonation, hence why traditional models like SLM and PAM need to be bolstered by additional intonation-specific models like the five surveyed above.

1.1.4 The role of transfer in L2 intonation

As discussed in §1.1.1, transfer must always be situated in a broader web of other inter-related factors, and L2 intonation is no exception. One reason is that L2 learners can be non-targetlike in ways that would not be predicted from transfer. For example, a learner may fail to attain nativelike values in terms of fine-grained details in the alignment of an F0 target even if the L1 and L2 are not systematically different in a way that would predict transfer to occur. He et al. (2011) explicitly make this claim in a study of focus marking in L1-Mandarin learners of L2
Dutch. The same case could be made for numerous other studies, such as Trofimovich & Baker's (2006) analyses of peak alignment in L1 Korean L2 English learners. In both cases, there is no clear evidence that L1 Mandarin or L1 Korean is the source of the divergence.

Another piece of evidence that transfer is not the whole picture comes from the fact that several studies have reported learners from typologically diverse L1s patterning similarly. Mennen et al. (2010) conduct a longitudinal study of the inventory, realization, and distribution of intonational categories in L1 Punjabi and L1 Italian learners of L2 English and finds striking similarities between the two groups. Further evidence comes from Hewings (1995), who examines tone choice in L2 English learners from three L1 backgrounds - Korean, Greek, and Indonesian. Here again, there is significant overlap between the learner groups, this time in the failure to use boundary rises for "socially integrative purposes" (e.g. contradiction, withholding agreement, or reservation). These two studies demonstrate that non-targetlikeness can result from one's simply being a learner (hence, e.g., lacking the requisite input in a sufficient range of discourse contexts) rather than one's L1 per se.

One final caveat regarding the role of transfer in L2 intonation concerns simultaneous multilinguals, for whom 'transfer' may be thinking about the mixing of two languages in the wrong way. For example, Queen (2012) argues that Turkish immigrants in Germany have drawn upon the intonation systems of their two languages to create a 'fused' grammar with a novel contrast between two kinds of rises. For these individuals, this is not a transient scourge that will later be shed at higher proficiencies; rather, it is a resource empowering them to make novel pragmatic nuances in conversation. It may also serve as an identity marker, especially for second- and third-generation immigrants. In addition to static interference (traditional L1-to-L2
transfer) and a fusion system (such as that described by Queen (2012)), Puri's (2013) study of Hindi-English bilinguals also identifies inherited influence as a third possibility, whereby some facet of one's parents' "nativized variety" (e.g. a World English like Indian English) is acquired as an L1. All of these examples simply serve to show how multilingual individuals may show more complicated patterns of linguistic intermixing of languages than the simple 'transfer' normally implies.

All of the above disclaimers notwithstanding, however, transfer is just as widely attested in the production and perception of L2 intonation as it is in other areas of segmental L2 phonology. This fact is evidenced by the numerous studies cited in Chapter 2, each exemplifying a different shape that intonational transfer can take. Moreover, Ulbrich (2013) finds that, as predicted by the construct of a 'shared phonetic space' in the SLM, transfer can 'backwash' from an L2 into one's L1 production. Thus, to the extent that transfer is ubiquitous and bidirectional, L2 intonation appears to pattern just like other aspects of L2 phonology.

1.2 Three components of the dissertation

This section lays out the three components - theoretical (§1.2.1), methodological (§1.2.2), and empirical (§1.2.3) - to the present dissertation. Each can be thought of as both a goal that the dissertation seeks to accomplish and, to the extent that goal is materialized, a research contribution. These three are inter-connected in that the theoretical component forms the framework for the empirical component, which in turn necessitates the methodological

---

5 Interestingly, this transfer can happen even in cases where, unlike segments, the relevant intonational categories are not marked in the orthography. This demonstrates that orthographically-induced abstract phonemic mappings are not a prerequisite for transfer to occur in L2 phonology.
1.2.1 Theoretical component

Section 1.1.4 established that, while numerous other non-transfer factors may be operative, transfer does in fact occur in L2 intonation, just like other aspects of L2 phonology. However, merely concluding that transfer occurs is far from the end of the story. Much is still ill-understood about exactly how transfer factors into the process of acquiring intonation in an L2. Three examples of questions still lacking satisfactory answers are: (1) What conditions are required for transfer to occur?, (2) How do transfer and non-transfer factors interplay?, and (3) How does proficiency level modulate when different kinds of transfer appear? The theoretical component of the present dissertation seeks to make progress in this direction, specifically addressing a fourth question: What is the range of possible shapes that L2 intonation transfer can take? While Mennen's (2015) LILt is a valuable first step towards thinking in this direction, much still remains to be worked out (especially since LILt is not focused on transfer specifically). For example, while LILt assigns a single dimension to "the inventory and distribution of categorical phonological elements", there are numerous ways that a cross-linguistic mismatch of this type could manifest itself in terms of transfer.

Toward this end, the present dissertation sketches out a typology, as referenced in the dissertation's main title, "Typologizing native language influence on intonation in a second language". The typology has two 'layers', i.e., it simultaneously achieves two ends. The first cross-linguistic layer establishes the space of the various logically possible ways in which the intonational phonology of two languages (acquired as native languages) can mismatch, loosely
based on the notion of Contrastive Analysis (Lado 1957). The second **transfer layer** applies the cross-linguistic layer to cases of L2 acquisition, spelling out the range of possible kinds of negative transfer in L2 learners' intonation production. The typology is structured into 6 'branches', each of which has 2 'nodes', for a grand total of twelve classes (of cross-linguistic mismatch and transfer).

The overall goal of spelling out a typology of this kind (not tied to any one pairing of languages) is to serve as a framework that future researchers can interpret their work with respect to, akin to the SLM. In a sense, what the typology seeks to do is similar to PAM L2 in that it is a typological collection, labeling various kinds of categories of crosslinguistic influence. The typology can be thought of as a production-oriented counterpart to PAM-S, since both models describe "i-Categories" and PAM-S focuses specifically on perception. Finally, the proposed typology fleshes out the implications of LILt for transfer. With its many nodes, its comparative and L2 layers, and its longitudinal component, the typology can be thought of as a finer-grained extension of the model.

**1.2.2 Methodological component**

A second piece of necessary background for the empirical component of the dissertation is a more refined way of handling L2 learners' F0 data. The predominant method for applying the AM framework to describe spontaneous speech is through transcription (ToBI or otherwise), i.e. by holistically marking time-stamped hypotheses about what phonological categories are believed to underlie the observed F0 track. This method is problematic in the context of analyzing L2 learners' intonation for several reasons. First, it is unclear what set of transcription labels one should draw on in the analysis. Using labels for the L2 categories alone (e.g. using
Mainstream American English ToBI for L2 English) commits the comparative fallacy (Bley-Vroman 1983) since it fails to allow for the possibility of L1 transfer in learners' interlanguage (Makino & Aoki 2012:93). Some authors have attempted to circumvent this problem by drawing upon the L1 and L2 systems in the transcription process (Anufryk 2012) or using 'narrow phonetic transcription' labels (e.g. 5 edge tone combinations plus 12 pitch accents in Gut (2012)). However, such approaches expand the space of possibilities so much as to create rampant ambiguity, hence a given stretch of an F0 track can often be transcribed numerous ways, leading to low inter-annotator agreement. Gut (2012:11), for example, finds a value of mean kappa value of 0.33 for the tone tier across five transcribers. Even in cases where a solution to the label-set problem is possible, it is still the case that transcription conflates the surface description of the observed shape of the F0 contour from the phonological interpretation thereof. This particular issue is not tied specifically to L2 intonation but rather applies to any kind of intonational transcription, even of native speaker data (i.e. the context where transcription frameworks like ToBI and RaP were originally developed).

In order to overcome this problem and make more explicit the mental model of an F0 contour that a researcher posits during the transcription process, a stylization of the F0 contour may be created.⁶ A stylization refers to a schematic representation of the F0 track that reduces the degrees of freedom in the data yet still captures the relevant information for an analysis. In order to capture the rich, linguistically-relevant information in the F0 track, the ideal stylization would meet three criteria. First, it would be target-based, i.e. capture the shape of the F0

---

⁶ The notion of 'stylization' as used in this dissertation is adapted from the idea of 'close-copy stylization' from chapter 2 of de Pijper (1983). See Hermes (2006) for a review of the various available approaches to stylization.
contour in terms of F0 targets (the turning points in the contour). Since F0 targets form the phonetic primitives of the AM framework, this would make the analysis directly relatable to a phonological model. Second, the ideal stylization would allow for the F0 movements between adjacent F0 targets to be nonlinear. Rates of glottal vibration do not change over time in perfectly straight lines, and the stylization should reflect this. Third, it should not be assumed a priori that all F0 movements conform to the same 'cookie-cutter' shape (e.g. a polynomial). Rather, the ideal stylization should flexibly capture the idiosyncratic shape of individual F0 movements.

Unfortunately, none of the existing approaches to stylization meet all of these criteria. For example, the Fujisaki model (Mixdorff 1998) does not meet the first criterion since it decomposes the F0 contour into a series of 'commands' that, in themselves, are not directly mappable onto F0 targets. The Dutch IPO approach to intonation analysis (de Pijper 1983) meets the first criteria but not the second, as it represents an F0 contour as a series of F0 targets connected by straight lines. Momel (Hirst & Espesser 1993) meets the first two criteria but not the third, as it uses quadratic splines for all F0 movements regardless of their shape. (The F0 targets are determined so as to minimize the error from this modeling assumption.)

Thus, the methodological component of the present dissertation documents a novel approach to stylization that meets all three of the above criteria. The method involves identifying the individual F0 targets that make up the F0 contour and storing them as [time, F0] points (Criteria #1). The nonlinear shape of the F0 movement between adjacent targets is then captured through an automatic best-fit algorithm (Criteria #2). Since the best fit is determined on a case-by-case basis, idiosyncrasies in the shape of particular movements are retained (Criteria #3). The
overall end-product is a method that converts an F0 track into a quantitative representation that captures natural F0 movement shape more realistically and more accurately. This can then be mapped onto phonological structure using targeted queries to extract out the frequency of specific phenomena of interest.

To the extent that the queries can map the stylized contour onto phonological categories, the proposed method can be thought of as an alternative methodological instantiation of the AM framework. Moreover, while other such AM-based methods discussed in §1.2.2 (ToBI, ToDI, IViE, and RaP) are all based on transcription, the proposed method is not. As such, it shares with the other stylization approaches the crucial separation of description and interpretation (here, the stylized contour vs. the queries thereof) that is lacking in transcription-based approaches. Thus, while the proposed method certainly suffers from its own share of problems (cf. §3.3.2), it is hoped that its distinct combination of attributes within the larger space of existing methods renders the method useful in future research on intonation.

1.2.3 Empirical component

Several authors have argued that functional transfer is more difficult and pervasive than other kinds of transfer. Toivanen (2003), for example, finds that twelve L1 Finnish L2 English learners acting a pre-written conversational dialogue could correctly produce both rising and non-rising nuclear contours at phrase boundaries. However, the circumstances in which they used the different kinds of nuclear contours was far from nativelike (in particular in "reserved" or "incomplete" statements). Further evidence comes from Atoye's (2005) test administered to 120 users of English in Nigeria on five intonational 'minimal pairs' (e.g. utterances with vs. without a pitch accent or tonally-marked boundary in a certain location). While 85.7% of responses on an
AX task (with all participants pooled) indicated that participants could auditorily detect the
difference between the pairs, only 25.7% could assign the intended pragmatic interpretation in a
free-response task. To the extent that the results from these two studies are attributable to L1
influence, they suggest that transfer of pragmatic function may be more prevalent than that of
phonetic form.

Another widely-reported finding is that non-targetlikeness in phonetic realization (i.e. the
exact alignment and scaling of some F0 target) is more frequent than non-targetlikeness at the
level of the phonological category as a whole. One example of a study that has made this claim is
Jun & Oh (2000). In their study, advanced L1 English L2 Korean learners could successfully use
phonologically distinctive phrase-final boundary tones to group words into phrases (as
modulated by the phrase's meaning and length). However, these advanced learners struggled to
produce the surface realization of other tones in a phrase (in ways that are not contrastive within
Korean). Again, to the extent that transfer is to blame for the non-targetlikeness, this suggests
that transfer at the level of phonetic realization may be more frequent than other more
'phonological' kinds of transfer. Several other studies reaching this conclusion are reviewed in
Mennen (2008).

The studies surveyed above constitute a valuable first step toward understanding, for any
given learner at any given point in time, what kinds of transfer are more vs. less frequent.
However, since they only exhaust a tiny fraction of the overall typology, we are unfortunately
still far away from a complete answer to this question. Moreover, it is also unclear how the
relative frequency of the different kinds of transfer is tied to proficiency (i.e. whether the
hierarchy of frequency is fixed across all proficiency levels or whether it changes as proficiency
increases). From a practical perspective, this means it is unclear which aspects of the L1 intonation system are more difficult for a learner to 'shed' from their L2 production. An investigation into this issue would help identify which kinds of transfer are transient markers of being a low-proficiency learner and which are retained even in high-proficiency learners.

The empirical component of the present dissertation seeks to shed light on these issues. It takes up three kinds of transfer phenomenon, each of which forms a different research question for the dissertation. In order to represent a broader diversity of kinds of transfer, these were drawn from nodes spread across different branches of the typology. The study then uses production data to determine how frequently these different kinds of transfer are observed in data from a set of learners with a shared L1 and L2. Accentedness/intelligibility rating data is then used as a proxy for proficiency to explore whether or not the same hierarchy of frequency obtains for beginning and advanced learners.

For solid evidence on the hierarchy of frequency, it is crucial to ensure the kinds of data used for each of the different kinds of transfer are directly comparable (e.g. all involving similar tasks). Towards this end, one particularly appropriate choice of data source is an internally homogeneous L2 speech corpus whose data all come from a single task (e.g. read-aloud elicited production). Another obvious advantage to using a corpus (as opposed to a more traditional kind

---

7 Keeping with the focus of the typology, the data used for this purpose are from learners' productions, i.e. inferring underlying mental structures based on patterns in the acoustic data. This, of course, only probes one half of learners' interlinked perception-production system. While perception is an important part of the picture, its role will not be directly addressed in this dissertation.

8 It is interesting to note that this 'comparability issue' is perhaps the reason why this kind of question is relatively rare in segmental L2 phonology. One particular challenge is that task difficulty (e.g. in choosing AX vs. sequence recall) is often conflated with construct difficulty (e.g. how inherently difficult phonotactics are to acquire in an L2 as opposed to some other aspect of phonology).
of psycholinguistic experiment) is that it makes it possible to explore the questions on a large scale, with sufficiently large N to factor out various sources of noise (such as that introduced by the arbitrary choice of stimulus materials). Another consideration is that the languages in question must be well-documented. Towards this end, American English and Tokyo Japanese were chosen, since the intonation systems of these two languages are among the best documented and best understood of the world's languages.9 Since these two languages were the first to receive comprehensive descriptions in the AM framework, they have formed the cornerstone of the framework throughout its development over the past 35 years.

The specific L2 speech corpus chosen for meeting all of the above criteria is the *English speech database Read by Japanese students*, or ERJ (Minematsu et al. 2002a). Thus, the present dissertation draws its data from L1 Japanese learners of L2 English - more specifically, native speakers of Japanese who are learning English as a foreign language (EFL) in Japan. This population will be referred to throughout the rest of this dissertation as "Japanese EFL learners", hence the sub-title of the dissertation ("Three transfer phenomena in Japanese EFL learners"). The three transfer phenomena under examination for this population are exemplified in the following table:

<table>
<thead>
<tr>
<th>Research Question</th>
<th>Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Catastróphic economic cutbacks neglect the poor.</td>
</tr>
<tr>
<td>2</td>
<td>{I cáught} {a strange insect} {in the inmost párt} {of the fórest}</td>
</tr>
<tr>
<td>3</td>
<td>Do you know this man in this photograph?</td>
</tr>
</tbody>
</table>

*Table 1.2: Examples of RQs*

---

9 For notational simplicity, 'American English' and 'Tokyo Japanese' will be shortened to simply 'English' and 'Japanese' throughout the rest of the dissertation.
For the first research question, a Japanese EFL learner may be expected to show an elbow on the second syllable (-ta-) of *catastrophic* due to influence from the phrasal H- in L1 Japanese. For the second research question, a learner may parse the utterance into many smaller phrases, demarcated with a low F0 target at the end of each phrase, due to the nature and frequency of the boundary L% in L1 Japanese. For the third research question, the boundary rise over *photograph* (indicating the utterance's status as a yes/no question) may be confined to the final syllable (-graph), due to the alignment of the beginning of boundary rises in L1 Japanese. The full motivation for each of these predictions is presented in Chapter 4.

### 1.3 Preview of the rest of the dissertation

The purpose of this chapter was to present background information on intonation and L2 acquisition (§1.1) and lay out the dissertation's theoretical, empirical, and methodological goals (§1.2). Having thus established these points and 'set the stage', the remainder of the dissertation is structured as follows:

**Chapter 2** details the proposed typology, citing example studies to illustrate the various kinds of L2 intonation transfer represented by its 6 branches and 12 nodes.

**Chapter 3** provides detailed documentation for novel methods to visualize F0 data as well as encode it as a representation storing not only [time, F0] points but also the shape of the F0 movement between adjacent points.

**Chapter 4** conducts an in-depth contrastive analysis in order to identify three critical differences between the intonational phonology of English and Japanese, thereby generating the three research questions for the present dissertation.
Chapter 5 details the process of selecting data from the chosen L2 speech corpus (ERJ) and the stages in processing it for analysis (including applying the method from Chapter 3).

Chapter 6 addresses the research questions from Chapter 4 by presenting the study's empirical results - not only of the intonational outcomes themselves but also the extent to which they are predicted by the ERJ's rating data.

Chapter 7 first interprets the results from Chapter 6 in terms of a hierarchy of frequency among the three types of transfer in question, then concludes the dissertation by identifying promising directions for future research.
Chapter 2: Typology

The typology proposed in this chapter can be thought of as having two 'layers'. The first cross-linguistic layer establishes the space of the various logically possible ways in which the intonational phonology of two languages (acquired as native languages) can mismatch, loosely based on the notion of Contrastive Analysis (Lado 1957). Identifying a difference between two languages first requires establishing some point of comparison. Accordingly, the typology takes as its starting six such points of comparison, each of which forms a 'branch' in the typology. These are listed in the following table.¹

<table>
<thead>
<tr>
<th>Section</th>
<th>Branch</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>§2.1</td>
<td>Position</td>
<td>the number of possible categories in a given prosodic position</td>
</tr>
<tr>
<td>§2.2</td>
<td>Category</td>
<td>inventory of categories in intonational phonology</td>
</tr>
<tr>
<td>§2.3</td>
<td>Accentuation</td>
<td>placement/distribution of pitch accents in an utterance</td>
</tr>
<tr>
<td>§2.4</td>
<td>Realization</td>
<td>phonetic realization of an otherwise analogous category</td>
</tr>
<tr>
<td>§2.5</td>
<td>Function</td>
<td>intonational expression of a pragmatic/discourse meaning</td>
</tr>
<tr>
<td>§2.6</td>
<td>Density</td>
<td>overall frequency of different kinds of intonational categories</td>
</tr>
</tbody>
</table>

Table 2.1: Names and descriptions for the six branches in the typology

¹ Note that the point of comparison represented by each of these six branches is a unitary entity like a category or a function, not a higher-order structure like a contrast or alternation. It should generally be possible to treat a case of the latter as an epiphenomenon and break it down into smaller pieces that fit into this typology.
Each of these six branches has two 'nodes', i.e. specific instances of the general class described by the branch. Information on each of the resulting twelve nodes is listed in the following table:

<table>
<thead>
<tr>
<th>Branch</th>
<th>Node (Section)</th>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Position</td>
<td>Emptiness</td>
<td>A given prosodic position (e.g. utterance-initial) can be phonologically unoccupied in one language but not the other.</td>
</tr>
<tr>
<td></td>
<td>Fixedness</td>
<td>A position must always be filled by the same category in one language; multiple categories are possible there in the other.</td>
</tr>
<tr>
<td>Category</td>
<td>Presence</td>
<td>Both languages can have tones in a prosodic position, but a category is present in one language and missing in the other.</td>
</tr>
<tr>
<td></td>
<td>Conditioning</td>
<td>Both languages have an analogous category but differ in the phonological environment in which it appears.</td>
</tr>
<tr>
<td>Accentuation</td>
<td>Flexibility</td>
<td>Pitch accents may flexibly occur in various locations in one language vs. always in one fixed location in the other.</td>
</tr>
<tr>
<td></td>
<td>Word types</td>
<td>Pitch accents are flexible in location for both languages but systematically associate with different kinds of words.</td>
</tr>
<tr>
<td>Realization</td>
<td>Alignment</td>
<td>An analogous F0 target in the two languages differs in terms of its temporal location relative to segmental landmarks.</td>
</tr>
<tr>
<td></td>
<td>Scaling</td>
<td>The presence-vs.-absence of some phonological process creates differences in the height of analogous F0 targets.</td>
</tr>
<tr>
<td>Function</td>
<td>Expression</td>
<td>Some specific pragmatic/discourse meaning is communicated through intonation in one language but not in the other.</td>
</tr>
<tr>
<td></td>
<td>Specific form</td>
<td>A given meaning is expressed with F0 in both languages but through different components of intonational structure.</td>
</tr>
<tr>
<td>Density</td>
<td>Pitch accents</td>
<td>The two languages differ in the total number of pitch accents that a single utterance generally contains.</td>
</tr>
<tr>
<td></td>
<td>Edge tones</td>
<td>The two languages differ in the number of edge (i.e. phrasal and boundary) tones that an utterance generally contains.</td>
</tr>
</tbody>
</table>

*Table 2.2: Description for each of the twelve nodes in the typology*

The second transfer layer applies the cross-linguistic layer to cases of L2 acquisition, spelling out the range of possible kinds of negative transfer in L2 learners' intonation production. In a sense, the typology can be thought of as defining twelve different dimensions along which
the construct of similarity in the Speech Learning Model (SLM) (Flege 1995) can be defined for the domain of intonation.

Four caveats are worth noting about the treatment of the transfer layer in the rest of this chapter. First, descriptions are cast entirely in terms of \textit{L1-to-L2} transfer. The exact same typology could be used to examine L2-to-L1 'reverse interference', but this is not the focus of the discussion. Second, the scope is restricted to negative, (i.e. detrimental) transfer only. Recall from §1.1.1 that negative transfer stems from \textit{differences} between languages - here, the various differences identified in the cross-linguistic layer. Third, the focus is on intonation production, not perception. While the effects of transfer are also frequently detectable in L2 learners' perception, in order to simplify the discussion, the description of each node is oriented from the standpoint of production. It is hypothesized that many aspects of this production-oriented typology also have direct parallels for perception as well, though this remains to be tested in future research. Fourth and finally, it is not claimed that all twelve categories are substantiated for any given L1-L2 pairing. Rather, the typology merely circumscribes the wider set of types from which individual instances of transfer are drawn from. In other words, the claim is merely that, when transfer is detectable in an L2 learner's intonation production, it is expected to take one (or more) of these twelve forms.

To create this typology, the available research studies on L2 intonation were collected and read through. The relevant conclusions were then extracted from each study in a standard format (roughly, \textit{L1=\_\_\_; L2=\_\_\_; Learners did \_\_\_\_; This may be transfer because \_\_\_\_\_\_\_\_\_\_}).\textsuperscript{2} These

\textsuperscript{2} As acknowledged in §1.1.4, transfer is only one of a larger host of factors influencing L2 intonation. As such, for many of the studies cited, alternative non-transfer explanations are often possible, some of which may have even been pointed out by the original author(s).
conclusions were then hierarchically organized and assembled into a larger structure. In this way, the structure of the typology was induced from the available empirical data in a bottom-up fashion.

One readily noticeable feature of the typology is the fact every branch has exactly two nodes. There no reason this has to be the case. Over time, as the typology evolves through being applied to more and mora data, this accidental feature of the typology will most likely disappear. At several points in the discussion in this chapter, potential room for expansion into additional nodes will be addressed.

The term 'analogous', referenced at several points in Table 2.2 and throughout the discussion in this chapter, is roughly equivalent to the SLM term 'similar'. While the transcription labels that may have been historically assigned to certain tones can guide thinking along these lines, similarity should be defined structurally at the level of the surface phonetic phenomenon. The emphasis is on the surface phonetic level because, from the perspective of the learner, an L1 and L2 structure may have a completely different underlying phonological source and still be treated as analogous.

In the discussion of the L2 implications for each typology node, two pieces of information are provided. First, the precise empirical prediction for the relevant kind of transfer is listed - i.e., if an L2 learner were exhibiting a certain kind of transfer, how it would be observable in data from naturalistic and/or laboratory tasks. Second, signaled by the keyword "over time...", the longitudinal learning challenge faced by the learner is laid out - i.e. what the learner gradually needs to do as their proficiency increases to cease exhibiting the relevant kind of transfer. For several of the nodes in the typology, these two pieces of information are broken
apart by L1-L2 *directionality*. In other words, for those nodes, given two languages 'A' and 'B', a finer-grained distinction is made between the kind of transfer that would be expected if Language A were the L1 and Language B were the L2 versus the converse case (where Language B is the L1 and Language A is the L2).

To attest to the fact that the various categories of transfer do in fact occur, each node is illustrated with several examples of published empirical studies showing evidence to that effect. The intention is merely to show that the various predicted kinds of transfer *can* occur, not how frequently they occur. Thus, while each proposed kind of transfer is well-supported by previous literature, only one or two studies are cited for each node, selected from the larger pool of studies as being convincing, representative examples of the type of transfer in question. Studies were also chosen so as to come from as diverse a range of L1-L2 pairings as possible.\(^3\)

In this way, this typology seeks to organize the existing research on the various manifestations of transfer in L2 intonation. As attested by the large number of studies cited in this chapter that were published just within the past few years, there has recently been an explosion of interest in the topic of L2 intonation. This makes it more important than ever to provide a common frame of reference as this field moves forward. It is hoped that the typology in this chapter can serve in this capacity.

---

\(^3\) In the discussion, studies on contact situations (e.g. World Englishes) are cited alongside ones from traditional classroom learning. This is based on the assumption that there is at least some overlap between the kinds of language mixing that occurs when two languages are in contact and when an adult learns an L2 in a classroom. While this assumption may or may not be correct, it will serve as a working hypothesis for the present dissertation. (See also the discussion in §1.1.4.)
The organization of this chapter directly mirrors the typology. There are six sections (§2.1 through §2.6), each corresponding to a branch in the typology. Within each section, there are two sub-sections (e.g. §2.1.1 and §2.1.2), corresponding to the two nodes within that branch.

2.1 Position

The first branch of the typology concerns cross-linguistic differences in the number of possible categories in a given prosodic position. This can be considered one sub-case of the systemic dimension in the L2 Intonation Learning theory (LILt), as it pertains to "the distribution of categorical phonological elements". At least five positions may be identified: utterance-initial, phrase-initial, phrase-internal (i.e. the pitch accent 'slot'), phrase-final, and utterance-final. These positions are akin to the construct of onset, nucleus, coda, and appendix in describing syllable structure. Note that edge tones account for four of these five positions. Depending on the version of the Prosodic Hierarchy (Selkirk 1978) one assumes, the notion of (prosodic) 'phrase' may be broken down further into more specific kinds of phrases (e.g., for Japanese, minor/accentual phrase vs. major/intermediate phrase). In addition, in language pairings where it is useful, additional lower-level positions may be defined (e.g. prosodic word initial/final).

First, §2.1.1 examines differences in whether a certain position can be phonologically 'empty' (i.e. not occupied by any category). Second, §2.1.2 examines differences in whether a certain position is 'fixed' such that it is only possible for one specific category to fill that slot. Note that these both essentially represent different points along a continuum of how many types of tone can occupy a given position, i.e. empty (N=0) < fixed (N=1) < variable/flexible (N=2+).
2.1.1 Emptiness

This node concerns cases where there can be no category in a given prosodic position (i.e. it can be 'empty') in one language but not the other. Two subtypes of 'emptiness' can be distinguished. First, the language could always have nothing in that position, e.g., how pitch accents are categorically absent from in Korean (Jun 2000). Alternatively, it might be merely possible for the language to have nothing in that position, e.g., how Japanese phrases can have no pitch accent if they are 'unaccented'. The second/other language (to which this is contrasted) can never have an empty position, e.g., how every prosodic phrase in English must have a pitch accent.

First, consider the case where a given position must be occupied in the L1 but can be empty in the L2. In this scenario, the learner may produce a category in that position even though that is ungrammatical in the L2. Over time, as the learner gains proficiency, they must learn to stop producing any categories in that position and instead leave it empty, as in the L2. Several examples of this kind of transfer have been attested in the literature. For example, Marković (2012:244) finds that Serbian EFL learners produce a %L at the left edge of words (even in pause-free running speech), just as in their L1 Serbian. Thus, yielding an utterance like "%L{ It seemed to take } %L{ an age to } %L{get there}". Similarly, Ng (2012:86-88) discusses the case of a word-final high tone (i.e. H%) in Chinese Colloquial Singapore English, which is argued to be transferred from Bazaar Malay creole. Thus, in the word 'responsibility' the H% would be associated with the final syllable (-ty), even though that is two syllables after the word's primary stress. In the English inventory (Beckman & Ayers-Elam 1997), there are no tones that attach at the level of the prosodic position 'word-initial' or 'word-final', hence those positions can be said
to be obligatorily 'empty'. Thus, both of these kinds of transfer exemplify one kind of 'emptiness mismatch'.

Now consider the opposite case, whereby a given position can be empty in the L1 but must be occupied in the L2. In this scenario, the learner may fail to produce a tone in the relevant position, i.e., they under-use the necessary tones there. Over time, the learner needs to begin producing the categories in that position as appropriate for the L2. This kind of transfer is also well-attested in the literature. For example, Mandarin lacks utterance-medial boundary tones, and L% and H% are the only two possible boundary tones at the end of an utterance (Peng et al. 2005:260). This stands in contrast to English, where L- and H- are both possible utterance-medial boundary tones, and where various utterance-final contours are possible - e.g. L-L%, L-H%, H-L%, and H-H% (Beckman & Ayers-Elam 1997). Not surprisingly, Xu (2009:52) finds that Mandarin EFL learners marked the vast majority (80%) of utterance-medial intonational phrase (IP) boundaries with a pause only (i.e. not tonally). Thus, the lack of phrasal tones in the L1 leads to their under-use in the L2, representing a second kind of 'emptiness mismatch'.

2.1.2 Fixedness

Languages can also differ in terms of the maximum number of categories that are possible in a given position. On one hand, some languages have a fixed pitch accent shape. For example, phonologically, Japanese pitch accents can only take the shape H*+L (Venditti 2005:179). On the other hand, other languages are more flexible and allow a wider variety of alternative pitch accent shapes. For example, the GrToBI standard for Greek describes 5 distinct pitch accents (Arvaniti & Baltazani (2005:86)).
Cross-linguistic differences along these lines have implications for transfer in L2 acquisition. In a scenario where the L1 is 'flexible' and the L2 is 'fixed', when speaking the L2, the learner may produce one or more of the L1 tones in that position. Over time, such learners need to constrain themselves to use just the one type from the L2 and stop tapping the richer L1 inventory. The opposite scenario is also possible - whereby the L1 is 'fixed' and the L2 is 'flexible'. In such cases, a learner may frequently use the L1 tone in their L2 production and, over time, needs to diversify their tone choice in that position.

Note that, in both of these scenarios, the exact outcome depends on whether the sole category from the 'fixed' language is a subset of the larger set of categories from the 'flexible' language. In the latter scenario, for example, if the L1 tone is not a member of the L2's set of tones in that position, then its use in the L2 is unconditionally ungrammatical. On the other hand, if the L1 tone is a member of the L2 tone set, then the non-targetlike aspect of the learner's production would instead be merely the fact that that tone is used disproportionately often (i.e. more so than native speakers).

An example of transfer stemming from a fixedness mismatch (of the 'L1=fixed, L2-flexible' variety) can be found in Puri (2013:81-96). Puri examines the intonation production of early and late L1-Hindi bilinguals of L2 English. Whereas English has a wide range of pitch accents, Hindi has a single rising contour, which has been variously analyzed as $L^*+H$ / $L+H^*$ or $L^*$ plus H%, among others (Puri 2013:42-46). In Puri's study, while early bilinguals produced a variety of pitch accents (e.g. $H^*$), as is targetlike for English, the early bilinguals produced only the Hindi rising contour on every non-final content word. Thus, the 'fixed' nature of tone choice
for this prosodic position in L1 Hindi seemingly transferred to L2 English production, thus constituting a case of transfer originating from a fixedness mismatch.

2.2 Category

The second branch of the typology involves comparisons at the level of individual intonational categories. This can be thought of as another sub-case of the systemic dimension of LILT in that it involves "the inventory of categorical phonological elements". First, §2.2.1 examines cases in which a category is present in one language but missing in the other. Second, §2.2.2 examines cases where a category is present in both languages but differs in terms of its phonological conditioning.

2.2.1 Presence

Here, both languages can have tones in a given prosodic position but one particular category is found in only one of the two languages. For example, both languages can have categories occupy the 'pitch accent' slot, but within the two languages' pitch accent inventories, one has L+H* whereas the other does not. This mismatch can be also mutual, such that in some given position, language A always has tone X and language B always has tone Y. Considering that languages vary widely in terms of their tonal inventories (Jun 2005, Jun 2014), it is not surprising that many pairings of languages exhibit differences in this regard. This kind of category-level mismatch is most directly in line with mainstream research in segmental L2 phonology (cf. how the missing category could be described as 'new' in terms of the Speech Learning Model).
Mismatches of this type can lead to transfer in both directions for a given L1-L2 pairing. First, in cases where a learner's L1 has the category but the L2 lacks it, a learner may use that category in their L2 production even though it is infelicitous in the L2. Over time, the learner must cease using that L1 category and instead choose from the L2 category set. Second, the converse case is also possible, whereby a learner's L1 lacks a certain category but the L2 has it. In such cases, a learner may not use that category in their production, and must learn to do so over time.

An example of the latter is provided in de Mareüil et al. (2015), which examines the variety of French spoken on the island of Corsica. (Parisian) French has an initial %L boundary tone at the beginning of an accentual phrase, but no %H (Jun & Fougeron 2000:214). In contrast, Corsican (another Romance language spoken on the island) does have an initial %H boundary tone, and this transfers into Corsican French (the "L2" in this case). For example, in the French sentence *La touriste trouve la cavité profonde* "The tourist finds the deep cavity", an F0 maximum may be observed on the first prosodic word (*la touriste*).

### 2.2.2 Conditioning

A second type of category mismatch involves cases where both languages have an analogous category but their phonological conditioning is different. 'Phonological conditioning' here refers to the environment in which the category appears, e.g. with only voiced or only voiceless obstruents. In cases where two languages differ in this dimension, a learner may use the phonological conditioning from the L1 when speaking the L2. Over time, they must shift

---

4 This is to be distinguished from a prosodic position, like utterance-initial or phrase-final, which is the subject of the 'Position' branch of the typology.
their use of the category in question towards being governed by the conditioning factors of the L2 rather than the L1.

One example of this kind of transfer is initial boundary tones in L1-Korean learners. In Seoul Korean (and other western varieties), the boundary tone at the beginning of an intonational phrase depends on the voicing of the phrase-initial consonant. If it is aspirated or tense, it is %H, otherwise it is %L (Jun 2000). English also has a %H, but it has no such link to consonant voicing (and is instead used to communicate a sense of contradiction) (Beckman & Ayers-Elam 1997:21). Similarly, Tokyo Japanese has an initial %L (Venditti 2005:180), but it is also unrelated to consonant voicing. This voicing-tone correspondence has been reported to transfer into the L2 production of L1-Korean learners of English. For example, Kim (2005:121) found that in the frame repeat __ again (with only two syllables before the target word), voiceless-initial words like team had a systematically higher F0 values across the word than voiced-initial words like dean, especially in low-proficiency learners. See also Konishi & Kondō (2011) for a similar finding in L2 Tokyo Japanese.

2.3 Accentuation

The third branch of the typology concerns the placement of pitch accents in an utterance.5 In contrast to the previous two branches, which were mostly concerned with the existence of intonational categories in a given language, this branch relates to the use or distribution of pitch accent categories in specific utterances. In other words, what is at issue here is where some kind

---

5 It possible to expand this branch to also account for the placement of edge tones in an utterance at prosodic breaks of various strengths. Diez (2008:331-332), for example, identifies a number of differences between Spanish and English in this regard. With this broader definition, the branch may be better labeled 'distribution' rather than 'accentuation'.
of pitch accent category is placed, not whether that category exists in a given language's inventory in a certain position. This can be thought of as yet another sub-case of the systemic dimension of LILT in that it deals with "the distribution of categorical phonological elements".

Mismatches relating to accentuation can come in two varieties. First, pitch accents may flexibly occur in various locations in one language but not the other (§2.3.1). Second, different kinds of words are pitch accented in the two languages (§2.3.2). Each of these two cases is discussed below in turn.

2.3.1 Flexibility

First, languages differ with respect to how flexible the location of pitch accents is. In some languages, for example, the location of pitch accents is more or less predetermined, e.g. "the final full (non-schwa) syllable of a prosodic phrase" for French (Fagyal et al. 2006:70). In contrast, in other languages (like English), the location of pitch accents reflects lexical stress locations, which vary from utterance to utterance. Moreover, in many languages, pitch accent locations may be even dynamically varied in order to communicate different meanings.

If accentuation is flexible in the L1 but fixed in the L2, the learner may produce pitch accents in the L2 at varying locations based on L1-mediated factors. Over time, such learners need to constrain themselves to produce pitch accents at only that one fixed location for the L2. In contrast, if accentuation is fixed in the L1 but flexible in the L2, the learner may consistently produce pitch accents at the fixed location from the L1 in their L2 production. Doing so may be non-targetlike given the lexical makeup of the utterance and/or discourse factors. Over time, the learner must learn to vary pitch accent location according to the L2 system.
As an example of the latter, consider the study by Horgues (2013). In certain prosodic positions (before utterance-medial boundaries and at utterance-final boundary rises), French EFL learners struggled with pitch accent location. For example, learners produced words in these positions such as *computér* and *protecțion* - with pitch accents in precisely the location dictated by the L1 (as described by Fagyal et al. (2006) above). Thus, the fixedness of pitch accent location is one way accentuation can transfer into L2 production.

### 2.3.2 Word types

For two languages that have flexible pitch accent locations, accentuation can differ in one other way: certain kinds of words could be accented in one language but not the other. Such differences can take many forms. For instance, a de-accenting process (e.g. reflecting focus or information structure) may apply in one language but not the other. Alternatively, certain categories of words (e.g. function words) may be categorically accented in one language but non-accented in the other. In light of such differences, a learners may place pitch accents in L2 utterances as dictated by their L1 system. Over time, the learner must acquire the L2 system governing the placement of pitch accents.

One example of this kind of transfer is discussed in Caspers et al. (2012), which examined L1 Hungarian and L1 German learners producing verum focus sentences in L2 Dutch. In such sentences, the negation word is pitch accented in Hungarian, whereas the verb is pitch accented in both German and the target language (Dutch). Caspers et al. (2012:33) found that nearly half of the utterances produced by L1-Hungarian learners had pitch accents on the negation word (or an adjacent adverb), as opposed to approximately 70% placement on the verb for L1-German learners, following the patterns from their respective L1s.
A second class of cases illustrating this node in the typology comes from whether Wh words are pitch-accented. In English, Wh-words are normally treated like other pronouns and are not pitch-accented (at least in Wh-questions produced without any contrastive focus). This is perhaps due to the fact that the status of an utterance as a Wh question is signaled syntactically, rendering a prosodic marking of that information less necessary. This stands in contrast to many other languages where, in the absence of a special kind of syntax specific to Wh-questions, the Wh item is indeed pitch accented. (See Kitagawa (2013) for relevant discussion.) Due to this difference, learners of English whose L1 is one such language may frequently place pitch accents on the Wh item in their L2 English utterances. This has been reported for L1s as diverse as Kannada and Bengali (Maxwell & Fletcher 2013), Russian (Crosby 2013), Japanese (Ueda & Saitō 2012), and Mandarin (Ji et al. 2012). For example, in *What does Mary know about Nelly?*, 'what' would receive a (L+)H* pitch accent in Kannada/Bengali English. In this way, the kinds of words that are pitch accented can differ from one language to the next, leading to transfer during L2 acquisition.

2.4 Realization

Another way that two languages can differ is in the phonetic realization of an otherwise analogous category, corresponding to the *realizational dimension* of LILt ("the phonetic implementation of categorical elements"). This kind of difference falls into two sub-classes: alignment (§2.4.1) and scaling (§2.4.2) - describing the horizontal and vertical dimensions of an F0 contour, respectively. Recall how, in the AM representational scheme, a tone category is phonologically specified for height and associated to the segmental string at some specific alignment. As such, these two dimensions account for the two primary ways that tones can differ
in realization.\textsuperscript{6} Since modulations in alignment and scaling generally create variation \textit{within} a given category (e.g. L\% or H\*), these often create L1-L2 differences that would be treated as 'similar' under the Speech Learning model.

\textbf{2.4.1 Alignment}

The first way that two languages can differ in terms of phonetic realization is in regard to \textit{alignment} - i.e. where a tone manifests itself in the time domain. In Autosegmental-Metrical phonology, alignment is treated as a reflection of which segmental landmarks a tone is phonologically associated ('time-locked') to, which can be as specific as an individual consonant or vowel (Ladd 2009). In some sense, alignment can be thought of as the intonational equivalent to fine-grained phonetic detail of a given phonological category in a narrow transcription.

The point of comparison for this node of the typology is some analogous F0 target in two languages, which may or may not come from the same underlying phonological source. For example, one may compare two languages in terms of the alignment of the low target for a L\* (e.g. the beginning vs. the end of a stressed syllable) - in other words, the phonetic detail behind a tone that is transcribed the same in the two languages. On the other hand, one may also compare two languages in terms of the alignment of a peak from a H*+L in one language but a H+L* in the other language. Due to this kind of discrepancy in the alignment of an analogous F0 target between a learner's L1 and L2, the learner may align it to the L1 landmark rather than the

\textsuperscript{6} In addition to alignment and scaling, it's possible that two languages could differ in terms of the shape of transitions (i.e. convexity vs. concavity) between adjacent F0 targets, hence this may transfer as well. Turco & Braun (2013) for example, find transition shape to be one of several ways in which boundary rises produced by L1 German L2 French learners are non-targetlike. However, as the study of transition shape is still in its infancy, it may be too early to tell if such effects are genuine cases of transfer \textit{per se}. Chapter 3 of the present dissertation describes an approach to quantifying transition shape, which can be of use in future research to clarify this issue.
L2 one. Over time, the learner needs to instead shift their alignment values to match the L2 landmark.

Mennen (1998) and Mennen (2004), for example, examine the case of the alignment of the peak created by a L+H* in L1 Dutch L2 Greek. Despite the fact that the relevant category is transcribed the same in the two languages, the peak is aligned systematically earlier in Dutch than in Greek. Thus, Mennen (1998) found Dutch-like early alignment of the peak in learners' L2 Greek productions. For example, in the Greek utterance *I paradhosi ton epiplon tha ginei tin Triti to proi* "The delivery of the furniture will take place on Tuesday morning", over half of the learners aligned the peak within the accented vowel of ‘parádhosi’, whereas the native speakers aligned the peak well into the following syllable (‘-dho-’). Interestingly, Mennen (2004) found that this also transfers from the L2 back to the L1, such that peak alignment in most learners' L1 Dutch was systematically later than Dutch native speakers with no experience learning Greek. 'Alignment transfer' of this sort is one of the best-attested phenomena in the L2 intonation literature. For other similar cases, see Colantoni & Gurlekian (2004) for L1 Italian L2 Spanish, Zięba-Plebankiewicz (2008) for L1 Polish L2 English, and Barnes & Michnowicz (2013) for L1 Venetian L2 Spanish.

2.4.2 Scaling

Another way languages can differ is in terms of the presence-vs.-absence of some phonological process that affects scaling, i.e. downstep/upstep or a global shifting of the F0

7 One methodological challenge in studies on L2 alignment is to make sure purported differences in alignment are not merely the byproduct of differences in rhythm or timing dynamics. In Yeou's (2010:71) comparison between native English speakers and Moroccan Arabic EFL learners, for example, what on the surface appears to be alignment differences actually was determined to be "due to syllable duration influence rather than to F0 alignment transfer from L1".
range up or down. By manipulating the F0 space itself, scaling processes of this sort would lead to differences in the realization of an otherwise analogous tone (e.g. a H*) in the two languages.8

In cases where the L1 does not have a certain scaling process but the L2 does, the learner may fail to modulate scaling in targetlike ways. Over time, the learner must begin applying the relevant L2 process. If, instead, the L1 has a scaling process that is absent in the L2, the learner may apply it to the L2, modulating scaling in the same ways as if it were the L1. Over time, they must learn to stop applying the L1 process and instead modulate scaling in L2-like ways.

One example of the latter kind of transfer is discussed in Maeda (2006). In Japanese Wh-questions, the Wh-word is focused, followed by a 'post-focal reduction', i.e. a compressed pitch range over the following few accentual/minor phrases depending on the meaning of the question (Deguchi & Kitagawa 2002; Ishihara 2003). This stands in contrast to English, where Wh-words are normally treated like other pronouns and are not pitch-accented (at least in Wh-questions produced without any contrastive focus), and there is no process analogous to post-focal reduction. Maeda (2006:74) found that post-focal reduction transfers into Japanese EFL learners' production of Wh questions like Why did you come to Japan?, such that the F0 range after the Wh word is systematically narrower in learners' productions relative to native speakers. This pattern was most notable in the group of learners without study abroad experience, but it could also be detected (albeit to an attenuated extent) even in the group with study abroad experience.

---

8 This node is only intended to capture transfer involving some phonological process affecting scaling, not a transfer of the overall level (F0 baseline) or span (F0 range) of entire utterances. While there is some evidence that such dimensions can vary from language to language (Mennen et al. 2014), it is exceedingly difficult to provide unambiguous evidence for transfer of this type. Such evidence would require factoring out numerous biological variables (age and sex), emotional variables (excitation), and phonological variables (presence of boundary rises). Moreover, even where that could be accomplished, it is possible that any observed transfer is paralinguistic and/or socio-cultural in nature, not linguistic per se. For all of these theoretical and methodological reasons, transfer of overall pitch level and span are excluded from the present typology.
Thus, in addition to alignment, processes affecting the scaling of F0 targets can also transfer from an L1 into an L2.

2.5 Function

All the branches in the typology up to this point have focused on intonational form, either in terms of position (§2.1), category (§2.2), accentuation (§2.3), or realization (§2.4). In contrast, the point of cross-linguistic comparison for this branch is some analogous function (i.e. pragmatic, discourse, or information-structural meaning) that is expressed intonationally in a given language. Since the point of departure is meaning, this branch corresponds to the semantic dimension of LILt, i.e. "the functionality of the categorical elements or tunes".

In the present usage, example 'functions' include contrastive focus, turn continuation, given vs. new information, utterance type (e.g. Wh question), and prominence relationships.9 Given the all-encompassing approach to the heterogeneity of intonational function, this branch of the typology could rightfully be split apart into several smaller branches.

Two kinds of mismatches along these lines are discussed below. First, a given function may be marked intonationally in one language but not the other (§2.5.1). Second, even if a function is marked intonationally in both languages, the exact way it is marked can differ (§2.5.2). Both of these can be seen as differences in form-function mapping, i.e. the relationship between 'what is being realized' and 'how it is realized'.

---

9 In theory, the use of F0 to signal lexical contrast could also be subsumed as a kind of 'function'. After all, such uses of F0 involve signaling to the listener some atom of meaning (the identity of a given word), much as intonation often signals information structure or other kinds of meaning.
2.5.1 Expression

The exact functions that are encoded intonationally varies widely from language to language. As such, it is a fairly common scenario for a given function to be marked intonationally in one language but not another. If a given function is marked intonationally in the L1 but not in the L2, the learner may nonetheless mark it in their L2 production even if that is unnecessary or redundant from the perspective of the L2. Over time, the learner needs to stop marking that function intonationally, and if the L2 marks that function by other means (e.g. syntax or a sentence-final particle), to do so instead. Conversely, if a function is not marked intonationally in the L1 but is in the L2, the learner may fail to use intonation to mark that function. If that function is expressed in other means in the L1 (e.g. syntax or a sentence-final particle), the learner may instead rely on those mechanisms to communicate the function in question. Over time, the learner needs to shed those other elements and transition to marking the relevant function through intonational means.

Hosseini (2013) provides an illustration of the latter kind of transfer. Hosseini discusses how contrastive focus is marked in Persian through non-intonational means (e.g. longer, louder, and more peripheral vowels). In contrast, Japanese uses a specific tone (H%) for this purpose, creating a "Prominence-Lending Rise", or PLR (Venditti et al. 2008:488ff). Hosseini succinctly summarizes the relevant empirical results as follows:

The second experiment examined L2 focused utterances and compared them with native speakers' utterances and demonstrated that PLR, which is present in Japanese but absent in Persian, is clearly observable in the speech of L1 Japanese/L2 Persian speakers as an L1 transfer effect in L2. On the other hand, most L1 Persian/L2 Japanese speakers fail to produce PLR in their focused constructions, due to its inexistence in their first language. (Hosseini 2013:66)
Hosseini uses the example of the Persian sentence *Mina ketāb-râ qarz gereft* "Mina borrowed the book", which L1 Japanese learners frequently produced with a superfluous boundary rise on 'ketāb-râ' (book-ACCUSATIVE). Conversely, the Japanese translation equivalent *Mina-ga hon-o karita* was frequently produced by L1-Persian learners without the normal boundary rise on 'hon-o (book-ACCUSATIVE).

This is by no means an isolated case. Another example of this kind of transfer is found in Gut et al. (2013), which examined the intonational marking of given vs. new information in Malaysian English. While the given-new distinction is marked intonationally in English through the de-accenting of given information, the same is not true for Malay, where it appears to be realized through non-prosodic means (Gut et al. 2013:189). Consequently, Gut et al. demonstrated that Malaysian English also does not mark given and new information with distinct pitch accent placement. Moreover, a perception experiment confirmed that utterance elements could not be unambiguously categorized according to their information status based on auditory information alone. Thus, this serves as a second example of how the presence vs. absence of an intonational marker for a given function (e.g. contrastive focus or information status) in the L1 can transfer into L2 production.

**2.5.2 Specific form**

Even if a function is marked intonationally in two languages, the exact way that it is marked can differ (e.g. through the use of a tone category in one language versus a scaling process in the other). A difference of this sort can be deterministic, such that one meaning is mapped onto form X in one language vs. form Y in another language). Alternatively, a difference can be more gradient in nature, such that one language uses mostly X and occasionally Y
whereas the other language using mostly Y and occasionally X. In L2 acquisition, to communicate the relevant function, a learner may use the intonational form from the L1 rather than the one from the L2. Over time, the learner must begin using the L2 intonational form instead of the L1 one to express that function.

One example of this kind of transfer is the marking of an utterance's status as a question in Mandarin Chinese and American English yes/no questions. In American English, this is generally marked intonationally with a boundary rise (i.e. to H%) and often with a change of the nuclear pitch accent to L*. In contrast, in pragmatically unmarked circumstances, yes/no questions in Mandarin (ending with the particle 吗 ma) normally do not have a boundary rise. Instead, an utterance's question status is signaled through a global upward-shifting of the F0 range for the utterance.

These various cross-linguistic differences create multiple manifestations of transfer for this language pairing. First, Viger (2007:66ff) demonstrates that L1 English learners of Mandarin do not globally shift their F0 ranges upward in Mandarin yes/no questions. Moreover, Pytlyk (2008) shows evidence that some learners transfer the H% from L1 English, thus producing infelicitous rises on the particle ma.

---

10 Mandarin does technically have a H% boundary tone that can be used on yes-no questions. Peng et al. (2005:248) give the example of  tāmen bù mài yúsān ma? 'Don’t they sell umbrellas?', whose meaning with an accompanied H% they describe as follows: "[T]he speaker is asking a yes-no question, but the boundary tone suggests a presupposition that the store should sell umbrellas. Thus, this can convey surprise, if the addressee is someone who was sent to buy an umbrella and came back empty-handed." Since the use of Mandarin H% is pragmatically marked in this way, its function can be taken as differing slightly from that of the English generic yes/no question.
Transfer has also been reported for the opposite directionality in this language pairing. L1 Mandarin learners of L2 English often fail to produce boundary rises on yes/no questions (Ji et al. 2009, Chen 2013). Even when learners do successfully produce a boundary rise, they often fail to change the nuclear pitch accent to L*, especially if it is multiple syllables away from the end of the utterance (Ji 2010:64). Taken together, these examples illustrate how, when a given function is encoded by different aspects of the intonation system in two languages, the differences can lead to transfer in both directions.

2.6 Density

The sixth and final branch of the typology concerns the overall frequency, or 'density', of different kinds of intonational categories, calculated relative to some linguistic unit (e.g. 'pitch accents per major phrase'). In a sense, this can be thought of as the overall global picture that emerges based on the finer-grained distinctions in form (branches 1-4) and function (branch 5) introduced above. This branch corresponds roughly to the frequency dimension of LILT, describing "the frequency of use of the categorical elements". Below, the discussion will be broken into two parts - density in pitch accents (§2.6.1) and density in edge tones (§2.6.2).11

2.6.1 Pitch accents

Languages vary widely in the density of pitch accents (i.e. 'head/center marking'). Some languages (most notably 'lexical tone' languages like Vietnamese or Thai) generally have a dense concentration of tones, each of which may be analyzed as special kind of complex pitch accent

---

11 This particular domain appears to be particularly susceptible to non-transfer (i.e. developmental or universal) processes. For example, pitch accent density has been attributed to hyperarticulation (Rosenberg & Hirschberg 2010), and edge tone density has been attributed to planning and cognitive load (Kaglik & de Mareüil 2010), and in read-aloud tasks, punctuation and syntactic parsing (Chen 2006).
(potentially also with syllable-level edge tones). Consequently, at an extreme, some utterances may have a pitch accent on every syllable (including those for function morphemes) in such languages. The opposite extreme of the continuum is exemplified by languages like Korean, which lack pitch accents entirely, thus making it possible for long phrases to have numerous unaccented syllables in sequence. English is an intermediate case in that every (minor) prosodic phrase must have a pitch accent but they are often spaced relatively far apart, e.g. with several intervening function words.

This typological diversity can lead to transfer in L2 acquisition. If the L1 has a sparse concentration of pitch accents, the learner may use too few of them (i.e. de-accent too much material) in the L2. Over time, such learners must begin to accent the appropriate words. In contrast, if the L1 is has a dense concentration of pitch accents, they may accent L2 words excessively, including words that normally would not be accented. Over time, the learner must begin use fewer pitch accents, i.e. de-accenting the appropriate material in order to space them further apart.

Cases of the latter are especially well-attested in the literature. Mixdorff & Ingram (2009), for instance, applied the Fujisaki model to elicited production data from 17 L1 Vietnamese L2 English learners, plus 3 Australian English native speakers as a baseline for comparison. Results suggested the frequency of accent commands (an indirect indicator of pitch accent density) was much higher in the learners' data: one command every 2.62 syllables for the native speakers as compared to every 4.47 syllables for the native speakers.

The L1 need not be a 'lexical tone' language for such transfer to occur. Hellmuth (2010), for example, obtains a similar finding with elicited production data from L1 Egyptian Arabic L2
English learners, using the prompts and map task from the Intonational Variation in English (IViE) speech corpus. Relative to the baseline (the same sentences produced by British English native speakers from the corpus), the learners produced pitch accents on a higher percentage of content words. The learners' high proficiency (as "advanced learners enrolled in English-medium postgraduate study") makes it unlikely the greater accent density is merely a byproduct of slower articulation. Hellmuth concludes that the most probable source of the over-accenting is transfer from L1 Arabic, a language where, unlike English, an accent routinely occurs on every content word. Thus, for certain language pairings, a denser concentration of pitch accents in the L1 can indeed transfer into an L2, both for tonal and non-tonal L1s alike.

2.6.2 Edge tones

The second kind of density mismatch involves the frequency of edge tones. Since edge tones demarcate prosodic constituents of various sizes, a difference in 'edge tone density' is simply an alternative framing of the fact that two languages may break an equal-sized utterance into different numbers of tonally-marked prosodic phrases. Such mismatches in the density of edge tones (i.e. 'edge marking') are often a reflex of the fact that some analogous tone is attached to some smaller ('minor') phrase in one language and a larger ('major') phrase in the other.

First, consider the case whereby the L1 generally has denser edge marking than the L2 (e.g. if minor phrases are tonally marked in the L1 but not the L2). In such cases, the number of tonally-marked phrase boundaries in a learner's utterance in the L2 may be excessively numerous. Over time, the learner needs to begin thinning out their use of edge tones in targetlike ways (e.g. by inserting them only at major phrase boundaries). The opposite case involves the L1 having sparser edge marking than the L2 (e.g. if the L1 only uses edge tones at major phrase
boundaries but the L1 does so for minor phrases). In such cases, a learner may produce utterances in the L2 with too few tonally-marked phrase boundaries. Over time, the learner must learn to increase their rate of edge marking in targetlike ways (e.g. by beginning to use edge tones at minor phrase boundaries).

One example of the latter kind of transfer is discussed in Vargas & Delais-Roussarie 2012. This study examined 35 L1 Mexican Spanish learners of L2 French performing three different tasks (reading aloud, monologues, and interview). Only 36.9% of the prosodic words produced by the learners had clear intonational boundary marking, as opposed to 82% for a comparison sample of 10 French native speakers. These results suggest a transfer of learners' spare edge tone density from L1 Spanish, where it is common to have longer (major) phrases with multiple pitch accents, unlike French.

The 6-branch, 12-node typology thus outlined constitutes the theoretical component of the present dissertation. The following chapter shifts the discussion to the methodological component, i.e. documentation of a novel approach to extracting, visualizing, and analyzing F0 data.
Chapter 3: Methodological framework

This chapter introduces the methodological framework adopted throughout the present dissertation. The discussion below is broken into four smaller pieces. First, Section 3.1 describes an approach to extracting and visualizing F0 data in a way that retains more of the rich information in the acoustic signal. Next, Section 3.2 introduces a new way of creating a stylization (i.e. a schematic representation) of an F0 contour. The proposed approach decomposes an F0 track into two components - vertices and transitions - which are treated in greater detail in Sections 3.3 and 3.4, respectively. Finally, section 3.5 documents a standard format for storing stylizations in a textfile and briefly describes how such data can be 'queried' to address a particular research question of interest.1

3.1 Adopted methods for F0 extraction and visualization

In the figures in this and subsequent chapters, F0 data is extracted and visualized in a somewhat non-traditional way. As such, a description of the methods used (as well as the rationale behind them) requires comment. The goal of this section is to clarify these issues, thereby establishing how F0 data will be handled throughout the rest of the present dissertation.

1 All code referenced in this chapter - most notably the functions ToPitch(), ReadPitch(), RichVisualization(), Stylize(), and PlotStylization() - are available online as an R package at https://github.com/usagi5886/intonation
First, section 3.1.1 motivates the choice of the F0 tracker selected for this dissertation. Section 3.1.2 then discusses the parameter settings that were applied when using the chosen F0 tracker. Final, section 3.1.3 presents a novel framework for the 'rich visualization' of F0 data.

### 3.1.1 Choice of F0 tracker

In order to make the results of the present dissertation more easily replicable and directly comparable to other published studies, the algorithm used to generate the raw F0 data from the soundfiles should be widely used and well documented. Of the ten algorithms listed in Tsanas et al. (2014), the following are the top three F0 extraction algorithms whose published documentation has been cited the most in the literature. (Citation counts are based on Google Scholar (http://scholar.google.com/) at time of writing (July 2015).)

<table>
<thead>
<tr>
<th>Name</th>
<th>Publication</th>
<th>Description</th>
<th>Citations</th>
</tr>
</thead>
<tbody>
<tr>
<td>YIN</td>
<td>de Cheveigné &amp; Kawahara 2002</td>
<td>Implemented as a function in Matlab; Code available at <a href="http://audition.ens.fr/adc/">http://audition.ens.fr/adc/</a></td>
<td>1,327</td>
</tr>
<tr>
<td>Praat (ac)</td>
<td>Boersma 1993</td>
<td>Autocorrelation-based algorithm; One of several F0 trackers available in Praat (Boersma 2001)</td>
<td>914</td>
</tr>
<tr>
<td>RAPT</td>
<td>Talkin 1995</td>
<td>Included in Snack Sound Toolkit (Sjölander 2004); Based on earlier get_f0() function in ESPS waves+</td>
<td>739</td>
</tr>
</tbody>
</table>

Table 3.1: Top three F0 extraction algorithms used most widely today

Of these three, YIN is used primarily in computer science, especially for digital signal processing of music (e.g. computational analysis of meter and melody for automatic music transcription). In contrast, the remaining two (Praat's autocorrelation method and RAPT [Robust Algorithm for Pitch Tracking]) are used frequently in speech sciences. This observation, combined with the citations counts above, confirm Evanini et al.'s (2011:8) claim that Praat's autocorrelation method is "by far the most commonly used method" in linguistics today. Since the present study is linguistic in orientation (being situated in the field of Second Language
Acquisition), the Praat autocorrelation algorithm is a natural choice for the purposes of the present dissertation under the criterion of wide usage.

A second (and perhaps much more important) criterion is that the F0 estimates be reliable and of high quality. In a recent systematic comparison of ten algorithms (including three in Praat) on synthesized version of the sustained vowel [a] (where the 'ground truth' F0 value is known \textit{a priori}), none of the three algorithms listed above were among the best performing ones (Tsanas et al. 2014). Instead, it was two entirely different algorithms - SWIPE' [Sawtooth Waveform Inspired Pitch Estimator prime] (Camacho & Harris 2008) and NDF [Nearly Defect Free] (Kawahara et al. 2005) - that were found to outperform all others. These results parallel similar findings reported in Camacho & Harris (2008). Thus, under this criterion, the Praat autocorrelation method is a less obvious choice.

However, Evanini et al. (2011:1) note that Camacho & Harris (2008) "used a fixed search range 40-800 Hz for all speakers, regardless of sex or speaker-specific pitch characteristics". The same is true for Tsanas et al. (2014), where the F0 range was fixed at 50-500 Hz. In their own comparison of five F0 trackers (three in Praat, SWIPE', and RAPT), Evanini et al. (2011) found that when the F0 range was fine-tuned for each individual soundfile, all five algorithms showed similar performance. Upon further analysis, Evanini et al. found that the superficially favorable results for algorithms like SWIPE' in previous studies were due to the fact that such algorithms are less susceptible to pitch halving errors, especially for speakers the use an overall higher F0 range (e.g. many females) for whom this kind of error is particularly frequent. Essentially, these results mean that the difference between the various algorithms covered above are very slight, only emerging in cases where an arbitrarily vague F0 range is inappropriate for the analysis of a
given file. Thus, Evanini et al. (2011) conclude that, so long as the F0 range parameters are carefully tuned, any of the algorithms covered above perform well.

Following Evanini et al. (2011), in all analyses throughout the present dissertation, F0 ranges are manually tuned for each individual file. Under these conditions, the above results suggest the F0 measurements obtained from Praat's autocorrelation algorithm should be of state-of-the-art quality. Moreover, being the most widely used F0 tracker in phonetics today, using this algorithm makes the results of the present dissertation more easily replicable and directly comparable to other published studies. Based on these criteria, the Praat autocorrelation algorithm was chosen for use in the present dissertation. When this algorithm is applied on a specified soundfile, a 'Pitch' object is created that contains various pieces of information concerning the F0 analysis. This kind of object can then be saved to a textfile representation, which can not only be opened by Praat but also be parsed by a script in most programming languages).

Since all analyses in the present dissertation are conducted in R (R Core Team 2014), the Praat autocorrelation algorithm was accessed (and the F0 extraction was performed) by using PraatR (Albin 2014) - a software architecture that makes it possible to control Praat with R code. More specifically, for each soundfile to be analyzed, a function call of the form `praat("To Pitch (ac)...", ...)` in R took a soundfile as input and created a Pitch object textfile on the computer's hard disk. To simplify the code, the PraatR command was wrapped in a convenience higher-level interface function `ToPitch()` to create the Pitch object textfile. A separate function `ReadPitch()` was also created for efficiently reading in the (long Praat format) Pitch object textfile into R as a matrix. By conducting the entire analysis entirely within R in this way,
without switching between two software programs (Praat and R), the data processing workflow was streamlined considerably.

### 3.1.2 Parameter settings

Praat's autocorrelation algorithm relies on a total of ten input parameters, listed below.\(^2\)

<table>
<thead>
<tr>
<th>Function</th>
<th>Parameter</th>
<th>Default value</th>
<th>In header when saved to textfile?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identifying possible F0 candidates for each frame</td>
<td>Time step(s)</td>
<td>auto</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Pitch floor (Hz)</td>
<td>75</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Maximum number of candidates</td>
<td>15</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Very accurate</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>Determining the optimal sequence of candidates for the entire file</td>
<td>Silence threshold</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Voicing threshold</td>
<td>0.45</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Octave cost</td>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Octave-jump cost</td>
<td>0.35</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Voiced/unvoiced cost</td>
<td>0.14</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pitch ceiling (Hz)</td>
<td>600(^3)</td>
<td>✓</td>
</tr>
</tbody>
</table>

*Table 3.2. Input parameters for Praat's autocorrelation-based F0 tracker ("To Pitch (ac)...")*

The default time step, 'auto', is 0.75 divided by the value for 'Pitch floor'. Thus, if the pitch floor is 75 Hz, the time step is 0.75/75 = 0.01 seconds (i.e. 10 milliseconds). The time step is yoked to the floor in this way in order to achieve a balance between processing speed (i.e. ensuring analyses do not take too long) and precision of the results (i.e. avoid under-sampling the contour). However, since all analyses and measurements for the present study are off-line, there

---

\(^2\) Since all soundfiles analyzed in the present dissertation are relatively short in length (containing roughly one sentence each), there is no need to set the F0 extraction parameters specifically for specific sub-parts of an utterance. Rather, a single set of parameters is used for each file.

\(^3\) 600 Hz is the default value for the ceiling when the 'Sound: To Pitch (ac)..' command is invoked from the button in Praat's Objects window. When performing F0 analyses in the editor window, the default is 500 Hz. It is unclear why separate defaults are used for these two interfaces to the same underlying function.
is no need to sacrifice the precision of the results for the sake of processing speed. Thus, the present study uses a fixed time step of 0.001, i.e., an F0 measurement is taken every millisecond (as if the floor was fixed at 750 Hz in the 'auto' default). This methodological decision has the added benefit of allowing all speakers' data to have equally fine-grained temporal resolution (unlike the 'auto' default, which yields more measurements for speakers with higher habitual F0 ranges, e.g. many females).

For each point in time ('frame') at which the acoustic signal is sampled, Praat's autocorrelation algorithm identifies several possible F0 values ('candidates') and ranks in order from most probable to most improbable. One candidate is always 0 Hz, a dummy indicator of the 'voiceless' candidate (e.g. indicating the unavailability of an F0 value in the middle of a silence). The 'maximum number of candidates' parameter represents how many total parameters to store per frame. The default is to store up to 15 candidates per frame. Since the number of candidates in the frames in an utterance follows a roughly Zipfian distribution (i.e., many frames with 1 candidate, many fewer with 2 candidates, slightly fewer with 3, and so on), curtailing the tail of this distribution at 15 throws away very little information. (This upper bound is most likely in place only to prevent the algorithm from recruiting dozens and dozens of spurious voiced candidates inside a voiceless/silent frame, thus inflating the size of the resulting Pitch object.) To err on the side of retaining all potentially useful information, the maximum number of candidates per frame was left at the default of 15 in the creation of all Pitch objects used in the present dissertation.

As mentioned above, for all analyses in the present dissertation, F0 ranges are set separately for each file. More specifically, the F0 range is manually adjusted to the values of the
lowest and highest reliable F0 points in a given soundfile (i.e. ones that are not noise), rounded to
the nearest 5 Hz unit. For example, if the lowest and highest reliable F0 points are 136.5 and
357.1 Hz, respectively, the range is set to 135-360 Hz.

Besides the time step, maximum number of candidates, F0 ceiling, and F0 floor, all other
'advanced' parameters are generally left at their defaults. This is done for two reasons. First, it is
impractical to manually adjust all ten different parameters separately for every file. Secondly, ad
hoc tweaking of the advanced parameters for individual files would mean this information would
need to be recorded and reported for every file in order for the study to be fully replicable. Thus,
leaving the advanced parameters at their defaults not only keeps the analysis feasible but also
removes the guesswork in interpreting the data from the dissertation.

In contrast, the time step, maximum number of candidates, and F0 ceiling are stored as
part of the textfile representation of each Pitch object, and the F0 floor can be inferred from the
lowest F0 estimate in the file. (See the rightmost column in Table 3.2 for which parameters are
stored in the textfile and which are not.) As such, these parameters can be safely adjusted on a
file-by-file basis without sacrificing replicability.

3.1.3 Visualization of F0 data

The normal way of visualizing F0 in Praat is with an image like the following. The
utterance being depicted is flap2 ("The pink carpeting") from the set of canonical files for
Mainstream American English ToBI.4

---

4 This image was produced by selecting the Pitch object generated from the flap2 soundfile plus corresponding
official TextGrid and selecting 'Speckle separately...' (indicating the words from the TextGrid should be plotted in a
Figure 3.1: One of the ways built into Praat for visualizing an F0 contour

This kind of representation is not specific to Praat; similar sorts of images are generated by all major acoustic-phonetic software packages. Consequently, this is the mainstream 'default' way of presenting F0 data in publications (e.g. journal articles) today. However, in this representation, all F0 values plotted the same (as a black dot), regardless of how reliable the F0 estimate is. Under certain circumstances this can make it difficult to separate the signal from the noise (pitch doubling/halving, segmental perturbations, background noise, artifacts of non-modal phonation, etc.) and identify what the 'underlying' contour is.

Figure 3.1 contains a case in which this problem becomes particularly dangerous. The F0 contour therein appears to contain a final rise of sizable magnitude. However, upon listening to this file, it is clear that this utterance is produced with plain declarative (i.e. 'final falling')
intonation, hence its official ToBI transcription as H* L- L%. The spurious rise is addressed on the miscellaneous tier of the official transcription with the note "creaky_so_rising", suggesting that some aspect of the random periodicity in this portion of creaky voice causes F0 trackers to hallucinate a nonexistent rise. This phenomenon has been documented before, cf. this quote from Gussenhoven (2004:9):

    Utterances ending in sonorant consonants (e.g. [m,r,l,w]) or vowels may end with a reversal of the F0 in the last part of the utterance, where the signal fades out, a phase that may be detectable to the pitch tracker, but is ignored in human perception. There appears to be no discussion in the literature, but it would seem reasonable to assume that it is due to a relaxation of the muscles controlling the frequency of vibration of the vocal folds.

    It is due to phenomena such as this that the official English ToBI guide (Beckman & Ayers-Elam 1997:14) includes the admonishment, "Transcribers must therefore learn when to trust their ears to catch such misparsings in the fundamental frequency track". However, since a researcher's own perception can be biased based on factors such as their native language and their linguistic experience, relying on it excessively is a somewhat less-than-desirable solution.

    The alternative solution proposed here is to retain the multidimensionality of the underlying F0 analysis more fully in the visual representation. This approach will be referred to as 'rich visualization' throughout the rest of the dissertation (since it involves enriching the visualization of F0). More specifically, in the present context, a larger subset of the information contained in a Pitch object textfile output from Praat's autocorrelation-based F0 tracker will be drawn upon in the creation of an F0 contour plot. Of the data contained in a Pitch object textfile, at least three pieces of information are available about each frame:

    1. the **time** point for this frame
    2. the (raw/unsmoothed) **intensity** of this frame, scaled from 0 and 1
    3. the total **number of candidates** considered in this frame
Likewise, at least three pieces of information are available about each candidate:

(2)   a. the **frequency** (i.e. F0) value associated with this candidate (N/A if voiceless)
      b. the 'strength' (i.e. goodness) of this candidate, scaled from 0 to 1 (N/A if voiceless)
      c. this candidate's **rank** (relative to the others in the same frame)

Of all this information, only the time and F0 point information, i.e. (1a) and (2a), is represented in the traditional visualization in Figure 3.1 above. However, all of the other information can be useful as well. In particular, each of the following can be thought of as (gradiently) indicating that an F0 estimate may not be reliable:

(3)   a. A frame that has low intensity (worsening the signal-to-noise ratio), cf. (1b)
      b. A frame that has many candidates (suggesting uncertainty in F0 estimation), cf. (1c)
      c. A candidate that has a low strength value (indicating low goodness), cf. (2b)
      d. A candidate that is ranked low (i.e. is dispreferred relative to other candidates), cf. (2c)

These kinds of information can be brought into the visualization in numerous potential ways. Out of a large space of logical possibilities, mapping intensity/(1b) onto point size and strength/(2b) onto (grayscale) color, as exemplified in the following figure, was found to be a particularly useful configuration. (All rich visualization plots in this dissertation are generated in R using the custom-built function `RichVisualization()`.)
Figure 3.2: Rich visualization (First-ranked candidates only)

In the figure, the top pane contains the raw waveform. The bottom pane contains two kinds of segmental information about the utterance: the words making up the utterance (and the boundaries between them) as well as a spectrogram (to assist in identifying where segments begin and end). The rich visualization of the F0 (and intensity) information is displayed (i.e. ‘supra-’ to) this segmental pane. The size of each F0 point is determined by the intensity value of that frame (multiplied by 1.5 to make the distinction between smaller and larger dots clearer). The darkness of each F0 point is assigned along a continuum between deep grey (grey50, #7F7F7F) for strength = 1.0 to off-white (grey95, #F2F2F2) for strength = 0.5 and below). Thus, large and/or deep grey points represent the most reliable F0 estimates. Conversely, small and/or faint grey points represent the least reliable F0 estimates.

In Figure 3.2, the spurious rise at the end of the utterance is made up entirely of small dots, indicating low intensity. Moreover, the first half of the spurious rise is faint in color,
indicating low strength values. Assuming the logic in (3), both of these observations suggest that this particular stretch of the contour represents unreliable F0 estimates. Moreover, almost no activity is visible in the spectrogram during this stretch of the contour, reinforcing this conclusion. In this way, the rich visualization achieves the original goal of clarifying where 'noisy' F0 estimates may be.

Rich visualization is useful for identifying segmental perturbations in the F0 track as well. The utterance in the above two figures is called 'flap2' because, in the final word *carpeting* [ˈkaɹ.pә.ɾɪŋ], a flap [ɾ] occurs in the middle of the large F0 drop, perturbing the F0 track. With the traditional representation in Figure 3.1, the researcher can only merely guess that this is indeed an artifact of the flap. Indeed, transcribers are explicitly trained to 'watch out' for cases such as these (Beckman & Ayers-Elam 1997:13-15). In the rich visualization in Figure 3.2, however, the F0 points over that stretch are significantly smaller and almost entirely faint in color, suggesting that portion of the F0 track is merely segmental perturbation. Here again, the rich visualization helps separate out the noise (and remove the guesswork) when visually interpreting an F0 track.

While Figure 3.2 is a significant improvement over Figure 3.1, they both share a problem of a different sort. In both, only the candidate that the Praat autocorrelation algorithm ranked first is plotted; all other candidates are left out of the visual representation. This is particularly problematic for frames where the voiceless candidate is ranked first, hence no F0 point is plotted. In a significant portion of cases, the best F0 estimate in these frames can be a potentially valuable source of information, e.g. indicating how and where the contour faintly drifts off into silence. Thus, by blindly trusting the algorithm's binary voiced/voiceless decisions, and by
plotting nothing for the frames thus determined to be voiceless, much potentially useful information is lost.

The solution adopted here is, for these otherwise blank 'voiceless' frames in the plot, plot the candidate ranked second (i.e. the most probable F0 value) and explicitly mark it as such. This approach is illustrated in the following figure, which is identical to Figure 3.2 except for the addition of these extra 'secondary' points.

![Figure 3.3: Rich visualization (Highest ranked voiced candidate in every frame)](image)

Each secondary candidate is plotted as a single black pixel, indicating the algorithm's 'best guess' for the F0 point in an otherwise voiceless frame. This salvages a series of valid F0 measurements around 700 ms, at the beginning of the closure for the [p] in 'carpeting'. While this also adds a bit of random noise, since noise points are faint in color and small in size, this does not detract from the overall interpretability of the plot.
In conclusion, a quick comparison of Figure 3.1 with Figure 3.3 makes it clear just how much more information is retained in the latter. It is worth emphasizing that both figures are based on the exact same 'Pitch object' textfile from the same F0 tracking algorithm; the only change between the two is how the data is being visualized. Since a richer set of information should help to make more sound inferences, this type of representation will be used in all plots and analyses throughout the rest of the dissertation.

It is important to note, however, that at least four different aspects of this visualization process are arbitrary. First, it is arbitrary what aspects of the signal are included in the visualization (especially since different F0 trackers output different kinds of information). Above, candidate strength and intensity are used, but the number of (voiced) candidates per frame, for example, could also be useful. Second, it is also arbitrary what visual dimensions are used. For example, the representation above does not recruit background color or point symbol shape/type in the visualization, but there is no reason these could not be used. Third, it is also arbitrary which aspect of the signal maps to which visual dimension. Above, strength is represented with grayscale color and intensity is represented with point size, but these mappings could just as well be switched. Finally, due to printing restrictions, the rich visualizations in the present dissertation use a grayscale color palette. However, in the absence of such restrictions, a full-color palette would greatly enhance the visualizations (e.g. by using a red-to-indigo continuum for primary candidates). Thus, the specific format of the rich visualizations used in this dissertation should not be thought of as a monolithic standard. Rather, the proposal is for is a flexible framework that can (and should) be used to highlight whichever aspects of the signal are useful in a given context for a specific research goal.
3.2 Stylization

As is the case with other kinds of production data, F0 measurements generally must be coded in linguistically meaningful ways before they can be analyzed systematically. Toward this end, in the methodological framework adopted in the present dissertation, the next step in the analysis is to describe the extracted F0 measurements with a *stylization*. Recall from section 1.2.3 in Chapter 1 that a stylization is a schematic model that reduces the degrees of freedom in the F0 data yet still captures the relevant information for an analysis. The approach to stylization proposed here involves taking the raw F0 data and boiling it down into two kinds of information: (1) the location of the turning points in the contour and (2) the shapes of the movements from one turning point to the next. Together, these two pieces of information can be used to reconstruct a simplified representation of the original F0 data.

The following two figures illustrate the stylization process. The following plot contains a rich visualization of a MAE_ToBI canonical file.
Figure 3.4: Third sentence inside MAE_ToBI canonical file 'made1', without stylization

Amid various kinds of noise, the above F0 track clearly has several peaks and valleys - the features of interest for an intonation analysis. The purpose of stylization is to quantify and extract out this information for analysis. From a signal detection theory standpoint, the goal is to maximize the hits (retain linguistically meaningful information) and correct rejections (exclude irrelevant noise) while minimizing the misses and false alarms.

The following plot contains a possible stylization of this F0 track. The 'underlying' F0 contour behind the noise is represented as a combination of white circles and curved lines. Inspection of the figure makes it clear that the stylization represents the raw F0 data rather well, with generally minimal divergence from the raw F0 data. Computationally, a stylization object is first created by using a function Stylize(). This stylization object is then passed to a separate function PlotStylization(), which superimposes the points and lines over the top of an existing F0 contour plot (rich visualization or otherwise).
The rest of this chapter goes into greater detail about how this process works. The discussion is broken into two parts, one for each component of the stylization: vertices in section 3.3, and transitions in section 3.4. Throughout the whole process, only the raw F0 data are used. No smoothing algorithms are applied, nor are frames with missing F0 values filled in ('imputed'). This is the case for two reasons. First, such 'pre-processing' is simply not necessary, i.e., the method has nothing to gain by modifying the raw data in these ways. While such procedures are normally taken to mitigate noise, the contour-fitting process discussed in §3.4.4 below already does so. Secondly, such algorithms always have the danger of distorting the data by creating methodological artifacts. For example, the specific smoothing algorithm and the degree of smoothing chosen, both of which are arbitrary, may impact the results (e.g. by throwing away relevant information along with the noise). For these reasons, no smoothing or imputation is applied to the F0 data at any stage in the analysis.

Figure 3.5: Third sentence inside MAE_ToBI canonical file 'made1', with stylization
3.3 Vertices

This section addresses two important issues regarding the notion of a vertex. The first issue is what linguistic unit vertices are intended to represent. Second, the relative merits of determining vertices automatically vs. manually will be discussed, as well as what sorts of criteria should be applied in the decision-making process.

3.3.1 What vertices represent

The core component of a stylization is the set of turning points in the contour (represented with white circles in Figure 3.5 above). Mathematically, each of these can be thought of as a point in the two-dimensional [time, F0] plane. In the present dissertation, a turning point of this nature will be referred to as a 'vertex' (plural 'vertices') - a term borrowed from the notion of a parabola's vertex (i.e. the 'inflection point' where it changes in direction). It is important to keep in mind that a vertex is merely a descriptive feature of the observed F0 track and does not necessarily represent the speaker's underlying gestural-phonological intention. This intuition is captured in the following analogy:

[Autosegmental-Metrical theory] assumes that intonation contours consist phonologically of strings of High and Low tones, which are phonetically realised as TONAL TARGETS, i.e. as specific points in the F0 contour, such as local minima and maxima. (This should not be taken to mean that local minima and maxima are equivalent to phonological tones, but only that the realisation of phonological tones gives rise to minima and maxima; this is comparable to saying that the second formant maximum in the word Maya is an important, easily measurable aspect of the phonetic realisation of the phoneme /j/, but it is not itself the phoneme /j/ and nor is it necessarily even the principal manifestation of this phoneme.) (Arvaniti & Ladd 2009:47) [Italics added]

In the present context, a vertex is merely the directly-observable 'specific point in the F0 contour' (e.g. a local minimum/maximum). In contrast, the term 'target' will be reserved for the speaker's underlying gestural-phonological intention (which is not directly observable). This
distinction becomes important in cases of undershoot, e.g. where multiple tones are crowded
onto one syllable. In such cases, the F0 change necessary to realize all of the targets would be too
rapid given the physiological constraints of the speech production system, hence the observed
vertices may end up being a poor reflection of the underlying targets. Such cases
notwithstanding, as Arvaniti & Ladd argue, phonological tones are generally realized in ways
that create corresponding vertices; hence, there is generally a rough correspondence between the
two.

3.3.2 Determining vertex locations

In the proposed framework, a researcher determines the locations of the vertices in a
given contour based on the same three kinds of criteria drawn on in intonational transcription: (1)
visual inspection, (2) perception, and (3) metalinguistic knowledge. The first criterion is visual
inspection of the contour, i.e., the researcher looks for where the F0 track changes direction (in
the case of a peak or valley) or changes velocity (in the case of an F0 'elbow'). This emphasis on
the actually observable contour-shape changes in the F0 track is something the proposed
framework shares with the Rhythm and Pitch (RaP) transcription system (Breen et al. 2012).
(Indeed, the approach to stylization documented in this chapter is loosely based on the pitch tier
of RaP.)

The second criterion is where the linguistically-relevant tonal movements begin and end
based on the researcher's listening to the utterance. This is a necessary component in order to
clarify what information in the F0 track is more or less perceptible in cases where the F0 track
contains misleading 'illusions'. Unfortunately, since this means the annotator's perception is
involved, this opens the possibility for the annotator's native language to bias the coding process. However, this is largely unavoidable and by no means particular to this framework.

The third and final criterion is the researcher's metalinguistic knowledge of the intonational phonology of the language being described. Recall from above that vertices are an indirect representation of underlying phonological targets. As such, a researcher's understanding of which tones are plausible in a given position (considering the segmental and prosodic structure of the text) is a valuable resource to tap. This means that the positing of a vertex is ultimately a phonological hypothesis, hence the stylization process is not language-neutral, in contrast to frameworks like Momel (Hirst et al. 2000). However, this way of thinking about stylization has the desirable advantage of leaving open the (likely) possibility that different features of a contour may be linguistically relevant for different languages. In particular, it is theoretically possible that a more-or-less identical F0 track could have two different sets of vertices in two different languages depending on what sort of segmental string it occurs over (prosodically structured as determined by the relevant language). By explicitly acknowledging that annotators should draw on their experience with the language in question during the coding process, this possibility is accounted for.

Since some aspects of the stylization process (especially the first criterion) involves quantifiable aspects of the F0 track, it could, in principle, be made fully automatic, and indeed this is an important task for future studies. However, before attempting to automate a given process, it is important to have a good idea of what the manually-generated gold standard should
be in the first place. With the method documented in this chapter still in its infancy, it was feared that a fully automatic approach may have been too premature, creating far too many errors (of a poorly-understood origin) to hand-correct. Thus, at least for the purposes of the present dissertation, the visual analysis of the F0 contour is performed manually. More specifically, the Stylize() function prompts the user to click on the desired [time, F0] points to be used as vertices. This means that, in the present dissertation, the present method is akin to transcription-based approaches (e.g. ToBI or RaP) in that both involve holistically marking hypotheses. However, the crucial difference is that the present method does so in a quantitative and phonetically explicit way.

3.4 Transitions

This section will address the issue of how the curves connecting adjacent vertices were modeled. First, linear and nonlinear approaches to stylization will be compared (§3.4.1). Then, the discussion will turn to the specific statistical distribution chosen to describe nonlinear transition shape (§3.4.2). The next section documents how that distribution can be re-parametrized more intuitively in terms of 'threshold' and 'gradience' (§3.4.3). Finally, the procedure for automatically fitting transition curves to actual F0 data will be introduced (§3.4.4).

3.4.1 Nonlinear vs. linear stylization

In many traditional approaches to intonational analysis (e.g. the IPO approach, cf. de Pijper (1983)), the vertices just described are the only piece of information stored about the F0

---

5 Note that this has parallels in intonation transcription: ToBI was first created for manual annotation in the early 1990s (Silverman et al. 1992) and then led to many attempts to automate various aspects of the transcription (cf. Rosenberg 2009 and references cited therein).
track, and the transitions between neighboring vertices are simply assumed to be straight lines. (In such a case, the vertices are said to be connected through 'linear interpolation'.) Under such an approach, the stylization would look something like the following.

![Figure 3.6](image)

*Figure 3.6: Third sentence inside MAE_ToBI canonical file 'made1', with linear interpolation*

A comparison of Figure 3.6 (assuming linear interpolation) with Figure 3.5 from earlier above (wherein nonlinearity is retained) makes it clear that a considerable amount of information is lost in the former (compared to the original signal), especially for the second transition. To facilitate comparison, the vertex locations have been standardized between the two stylizations. To be fair, a linear stylization would be able to achieve a slightly better fit than the one presented here if this constraint were not in place. However, regardless of where the vertices are placed, it is impossible for the the linear stylization to fully approximate the superior fit of the nonlinear stylization (without increasing the number of vertices in a *post hoc* manner).
The comparative performance of these two kinds of stylizations can be quantified in several ways. For example, one approach is, for every frame within a range of time, to subtract the raw F0 value minus the stylized F0 value, take the absolute value of the result, and then calculate the median across all the frames. Such a measurement (called the "median absolute deviation", or MAD) indicates the most typical size of the divergence between the raw data and the stylization model, regardless of sign (+ or -). Using this criterion, the nonlinear stylization has a MAD of 1.565 Hz, whereas the linear stylization's MAD is 10.002 Hz. In other words, the goodness of fit is over six times worse for the linear approach, suggesting strongly that F0 trajectory shape is pervasively nonlinear in nature. Using Bååth (2012), an implementation of the Bayesian equivalent of a two-sample t-test, to compare the MADs for these two stylizations of this particular utterance, after 20 thousand burn-in steps and 20 thousand samples, the mean improvement of the nonlinear model over the linear model is 6.36 Hz and the 95% Highest Density Interval (HDI) is 5.75 to 6.91 Hz. The HDI indicates that there is a 95% chance that the underlying mean generating the distribution (of the extent of improvement in Hertz) is between these two values. Crucially, this range excludes zero, hence it can be concluded that the nonlinear stylization retains substantially more information than the linear stylization. While this result is of course tied to this specific utterance, this pattern is ubiquitous - a nonlinear stylization can almost always be shown to be a significantly better representation of the raw F0 data.

This difference is not merely a matter of principle. The shape of pitch movement has been shown in several studies to systematically impact the interpretation of an utterance in a

---

6 Not surprisingly, this difference is also highly significant in a standard t-test.
variety of languages. (See, for example, Petrone & D’Imperio (2008) for Italian, Kaiser & Baumann (2013) and Dombrowski (2013) for German, Barnes et al. (2010) and Barnes et al. (2012) for English.) Thus, storing only the vertices would necessarily mean throwing away linguistically relevant information in the signal. For this reason, in addition to the vertices themselves, the nonlinear shapes of the F0 movements connecting neighboring vertices (as illustrated by the curved lines in Figure 3.5) are also quantified and stored for all analyses in the present dissertation. In articulatory terms, this can be thought of as meaning that not only are the (observable manifestations of the) gestural targets stored, but also the trajectory of movement for each of those gestures.

3.4.2 Distribution chosen for describing transition shape

The mathematical details behind how transition shape is quantified for the present dissertation will now be described in detail. The cornerstone of the proposed framework is the beta distribution, a statistical distribution described by two parameters, sometimes called 'shape 1' and 'shape 2' but referred to throughout the present discussion as 'A' and 'B'. Figure 3.7 below illustrates how these two parameters trade off.
Figure 3.7: Shapes of the beta distribution for values of A and B between 1 and 5

Each cell in the 5x5 figure represents a combination of A and B. For clarity of presentation, the values for A and B are integer values between 1 and 5, but numbers less than 1 and decimal values are also possible. The first number listed at the top of each cell is A, and the second is B. The axes are standardized across all plots. The x axis (extending from 0 to 1) represents some arbitrary length; hence, e.g., x=0.75 can be thought of as 75% across this length. The y axis represents the relative probability density over any given point on the x axis. The
The scale of the y axis (here, extending from 0 to 5) is essentially arbitrary - what matters is the shape of the distribution defined along the y axis between x=0 and x=1.

A few points are worthy of mention concerning this distribution:

1. If A and B are both 1, the result is a straight line.
2. If B=1 (top row), then the function is monotonically (i.e. unilaterally) increasing. Conversely, if A=1 (leftmost column), then the function is monotonically decreasing.
3. The two parameters are symmetrical, such that switching A and B has the effect of horizontally reversing the function. (Compare, for example, [5,1] and [1,5].)
4. Across the diagonal (from top-left to bottom-right), the curve becomes more and more centrally concentrated (and the tails of the distribution become thinner and thinner).
5. Generally speaking, A can be thought of as 'pulling' the distribution toward the right, and B can be thought of doing so towards the left. Thus, the center of the distribution is at exactly 0.5 when these two 'forces' are equal (e.g. at [5,5]), it is far to the right when A is larger than B (e.g. [5,2]), and it is far to the left when B is larger than A (e.g. [2,5]).

Of central interest in the present context is actually not the beta distribution itself but rather its cumulative distribution function (the pbeta() function in R), referred to henceforth as the 'cumulative beta distribution'. These functions are plotted below for the same combinations of A and B as in Figure 3.7 above.
Simply put, for a given value along the x axis, a cumulative distribution function describes what percent of the original distribution (here, the beta distribution) has been reached up to that point (working from the left to the right). Accordingly, every function is monotonically increasing, and y axis extends from 0 to 1 (i.e. 0% to 100%). Note how [1,1] is a simple straight line from {0,0} to {1,1} - indicating that, as one proceeds from the left to the right of the top left cell in Figure 3.7, the percentage reached across the original distribution increases linearly. All others cells in Figure 3.8 have some degree of curvature, representing the curvature in the

Figure 3.8: Shapes of the cumulative beta distribution for values of A and B between 1 and 5
corresponding functions in Figure 3.7. For example, the point at which the curve is highest in
Figure 3.7 corresponds to the point of greatest velocity in Figure 3.8.

The cumulative beta distribution, as illustrated in Figure 3.8, is ideal for modeling F0
transition shape. Since almost any combination of A and B is legal, it can successfully model a
multitude of possible transition shapes, the majority of which represent a physiologically natural
'S'-shaped gestural trajectory. While all of the transitions in Figure 3.8 would represent an
increase in F0 from one vertex to the next, all that needs to be done to model a decrease in F0 is
to vertically 'flip-flop' the curve. Thus, with just the two parameters A and B, both rising and
falling transitions of numerous shapes can be successfully captured.

3.4.3 Re-parametrization in terms of threshold and gradience

There are two disadvantages to using the cumulative beta distribution as-is. First, A and
B can in theory be less than 1, where the behavior of the function is rather different from that
described above (and the natural 'S' shape is lost). Second, the two parameters A and B lack full
mathematical independence in two ways: (1) the location of maximum velocity is reached
depends on the ratio between A and B, and (2) increasing the steepness of the transition (going
from the top-left to the bottom-right cell) requires increasing both A and B in tandem. To rectify
these inelegances, A and B are re-parametrized in terms of 'threshold' (t) and 'gradience' g as
follows.\textsuperscript{7}

\begin{align}
\text{(4) \hspace{1cm}} & \quad \text{A} = (2 \times t) \times g + 1 \\
\text{a.} & \hspace{1cm} \text{B} = (2 \times (1-t)) \times g + 1
\end{align}

\textsuperscript{7} Note that this re-parametrization was not adapted from the literature but rather constitutes a novel proposal devised
for the purposes of the present dissertation. The motivation for why these formulas look the way that they do will be
presented shortly below.
Through simple arithmetic, these equations can be solved for threshold and gradience as follows:

\[(5)\]
\[a. \quad t = \frac{(A-1)}{(A+B-2)}\]
\[b. \quad g = \left( \frac{(A+B)}{2} \right) - 1\]

For reference, the following table shows how the original A/B parametrization maps onto the new threshold/gradience parametrization in terms of the same combinations of A and B as Figure 3.7 and Figure 3.8 above. (In each cell, the first number is threshold and the second number is gradience.)

<table>
<thead>
<tr>
<th>A</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>[0.50, 0.0]</td>
<td>[1.00, 0.5]</td>
<td>[1.00, 1.0]</td>
<td>[1.00, 1.5]</td>
<td>[1.00, 2.0]</td>
</tr>
<tr>
<td>2</td>
<td>[0.00, 0.5]</td>
<td>[0.50, 1.0]</td>
<td>[0.67, 1.5]</td>
<td>[0.75, 2.0]</td>
<td>[0.80, 2.5]</td>
</tr>
<tr>
<td>3</td>
<td>[0.00, 1.0]</td>
<td>[0.33, 1.5]</td>
<td>[0.50, 2.0]</td>
<td>[0.60, 2.5]</td>
<td>[0.67, 3.0]</td>
</tr>
<tr>
<td>4</td>
<td>[0.00, 1.5]</td>
<td>[0.25, 2.0]</td>
<td>[0.40, 2.5]</td>
<td>[0.50, 3.0]</td>
<td>[0.57, 3.5]</td>
</tr>
<tr>
<td>5</td>
<td>[0.00, 2.0]</td>
<td>[0.20, 2.5]</td>
<td>[0.33, 3.0]</td>
<td>[0.43, 3.5]</td>
<td>[0.50, 4.0]</td>
</tr>
</tbody>
</table>

Table 3.3: Values of A and B from 1 to 5 mapped onto the new threshold (t) and gradience (g) parameters, in the format [t, g]

The new pair of parameters can be thought of as effectively rotating the 'axes' of the table, with threshold representing the bottom-left to top-right axis and gradience representing the top-left to bottom-right axis. First, consider the first parameter (threshold, t), which directly expresses the ratio between A and B as a number between 0 and 1. Along the diagonal (bolded in the table), where A and B are equal, the threshold is 0.5.\(^8\) Where A is bigger than B (top-right half of table), it shifts towards 1, and when B is bigger than A (bottom-left half of table), it shifts

---

\(^8\) Technically, in the case of \([A=1, B=1]\) (or anywhere gradience is 0), the value for threshold is undefined (because it incurs division by zero). However, it can easily be shown that the threshold asymptotically approaches 0.5 in this case (e.g. by calculating the threshold with A and B set to something like 1.0001 instead of exactly 1).
toward 0. The greater the imbalance between A and B is, the closer toward 0 or 1 the threshold becomes. Now consider the second parameter (gradience, $g$), which expresses the collective magnitude of A and B together. Note how an increase of +1 in A or in B each independently increases the gradience by 0.5. This parameter can be anything from 0 to infinity; that is, unlike the threshold, this parameter has no upper bound. Thus, for example $A=101$ and $B=101$ would cause the gradience to be 100.

The advantage of reparametrizing the distribution in this way is that the parameters are not only mathematically independent but also transparently map onto the shape of the contour. The following figure illustrates what happens to the contour if one of these parameters is held fixed and the other is free to vary. In the left panel, the gradience is fixed at 5 and the threshold is set to steps of 0.1 between 0 and 1 (i.e. 0.1, 0.2, 0.3, ..., 0.9, 1.0). In the right panel, the threshold is fixed at 0.5 and the gradience is set to whole numbers from 1 to 10. The color scale used in both panels is such that low values of the varied parameter (e.g. 1 and 0.1) are black and high values (e.g. 9 and 0.9) are faint grey.
Figure 3.9: Illustration of how modulations in threshold and gradience affect contour shape

The threshold modulates the location of the point of maximum velocity in a given contour (marked with a dot for each line in both panels of the figure). Thus, in the left panel, it can be seen that as the value of threshold increases (from black to faint grey), this point moves from the left to the right, which has the effect of shifting the temporal dynamics of the entire contour to the right as well. Thus, this point can be thought of as acting as the 'threshold' for the F0 change, effectively 'sliding' the contour left or right. The gradience parameter modulates how quickly the transition from y=0 to y=1 occurs (centered around the threshold). The lower the gradience value is, the more gradual and slow the transition becomes (cf. the blackest, straightest line in the figure). At the limit (i.e. gradience=0), the transition is a simple straight line. Conversely, the higher the gradience value is, the steeper the slope of the transition becomes (cf. the faintest grey line in the figure). At the limit (i.e. gradience=infinty), the transition would abruptly 'jump' from y=0 to y=1 at x=0.5. Thus, the gradience parameter describes how gentle or sharp the F0 change is.
3.4.4 Automatically fitting transition curves to actual F0 data

In the proposed framework, the parameters describing the shape of a transition are determined automatically based on the actually observed F0 data. This distinguishes this framework from the Momel algorithm (Hirst et al. 2000), where all transitions are assumed *a priori* to conform to the same quadratic shape (and the vertices are chosen in order to minimize the error due to this modeling assumption). More specifically, the present framework involves fitting the cumulative beta distribution (as reparametrized above) to the actually observed F0 points between each pair of vertices. For each transition, it is determined which combination of parameters (threshold and gradience) creates the most accurate representation of the raw F0 track. The rest of this section describes in greater detail how this fitting process works.

Barnes et al. (2014:1126) found that F0 information from more sonorous regions (in their case, [j]>[n]>[v]) are "accorded heavier weights" and therefore "extend relatively greater influence over" the perceived scaling of an F0 event. Acoustically, the result is likely due to differences in spectral cues (e.g. the richness of harmonic structure) and intensity over the consonants carrying the F0 information. Thus, while 'best fitting' was defined in section 3.4.1 above as simply the median absolute deviation (MAD) of the stylized F0 from the actually observed F0, these results suggest that an intensity-weighted version of this measure would be more perceptually realistic. Accordingly, the absolute deviations are weighted by intensity before calculating the median, as described in the following formula (in pseudo-code):

\[
(6) \quad \text{Fit} = \text{Median}(\text{Intensity} \times |\text{ObservedF0} - \text{StylizedF0}|)
\]

As before, for every (voiced) frame within the range of time for a given transition, the difference between the observed F0 value and the stylized F0 value is calculated, and the
absolute value of the result is taken. Next, this value is multiplied by the raw (i.e. un-smoothed) intensity for that frame (as stored in the Pitch object). Such values are scaled from 0 to 1, hence a deviation in a loud frame will be weighted heavily whereas an equally-sized deviation in a quiet frame will be attenuated in magnitude and therefore exert minimal influence. Finally, the median is calculated over all voiced frames within the relevant range of time. Any voiceless frames (hence 'secondary candidates' therein) are excluded from the calculation of the median.

Three desiderata are worthy of mention regarding the above formula. First, since the deviations between the observed F0 values and the stylization are weighted by intensity, the resulting values are no longer defined in Hertz but rather constitute their own unit. Second, since the units are derived from Hertz, any consequent nonlinearities (e.g. greater deviation values due to being in a higher pitch range vs. a lower one) are not addressed. Finally, the above formula does not take into account the number of frames going into the calculation of the median (e.g. the potential for the intensity-weighted MAD to increase due to 1000 frames going into the calculation as opposed to 100). Crucially, these last two issues are only relevant for comparing MADs across different transitions. During the process of comparing different possible transition shapes for a given range of time, the pitch range and number of points will be the same, hence the calculation of best fit is not compromised.

The ultimate goal of the curve-fitting process is to identify the transition shape that produces the smallest value for the intensity-weighted MAD (henceforth simply 'weighted MAD') as defined above. To do so, as a first-pass approximation, the weighted MAD is calculated for all 100 logical combinations of 10 threshold values and 10 gradience values. Collectively, these can be thought of as creating a coarse 10x10 grid across the 'search space'.
More specifically, the threshold values are 10 equal-sized steps from 0.01 to 0.99, and gradience values are 10 equal-sized steps from 0.01 to 5. Rounded to three digits, these are the following:

\[\text{(7) a. Threshold } = 0.010, 0.119, 0.228, 0.337, 0.446, 0.554, 0.663, 0.772, 0.881, 0.990\]
\[\text{b. Gradience } = 0.010, 0.564, 1.119, 1.673, 2.228, 2.782, 3.337, 3.891, 4.446, 5.000\]

If used to plot rising F0 transitions, these parameters look like the following. In the figure, color represents threshold (low/left=black, high/right=grey) and line thickness represents gradience (low/gentle=thin, high/abrupt=thick). Since the 100 contours largely fill the square box for the plot and represent a wide variety of shapes, this first-pass approximation does a reasonably good job of exploring the search space.

![Figure 3.10: Transition shapes as defined by the 100 combinations of 10 threshold and 10 gradience values](image)

All 100 contours are then compared with the raw F0 data, and each's weighted MAD is calculated. The 100 combinations of threshold and gradience parameters can then be ranked in
terms of goodness of fit, and the top-ranking parameter combinations are retained for subsequent calculations. Exactly how many top-ranking parameter combinations are used is a free parameter that can be specified in the `Stylize()` function (via the nSeeds argument). For the purpose of the present dissertation, the default nSeeds value of 10 was used, i.e. the top-ranking 10% of the 100 parameter combinations were retained for subsequent calculations.

Each of these 10 top-ranking parameter combinations is then used to initialize the `optim()` ('general-purpose optimization') function in R. Each parameter combination can be thought of as defining a 'seed' in the search space from which the algorithm explores similar parameter values in an effort to minimize the weighted MAD (i.e. improve the fit). Using the 10x10 grid to empirically determine ideal locations for these seeds helps to avoid artifacts from initializing the search in an inappropriate region of the search space. The following heat plot illustrates the nonlinear optimization problem involved in exploring the search space. The example used therein is the second-to-last transition in Figure 3.5, i.e. the F0 rise over made the to the peak on mar(malade).

---

9 A suitable optimization method in the present context needs to not only allow for two-dimensional optimization (for both threshold and gradience) but also allow for bounds (or 'box constraints') on the search space. In particular, the method must allow for a lower bound of 0 for both parameters and an upper bound of 1 for the threshold parameter. Of the various available methods for the `optim()` function, "L-BFGS-B" ("Limited memory - Broyden/Fletcher/Goldfarb/Shanno - Boxed") meets both of these criteria, hence it was selected for the present purposes. See the R help page for `optim()` for references and further information about this algorithm.
Figure 3.11: Topography of weighted MADs across many parameter combinations, and areas explored by the 10 seeds

In the plot, the x axis represents threshold and the y axis represents gradience. The color of each pixel making up the plot (the 'z' axis) represents the weighted MAD, with black representing the worst fit and grey representing the best fit. The grid superimposed over the plot represents the 10x10 grid of parameter combinations used in the initialization process (as described above). To the plot, 10 lines are added to represent the 10 seeds, each extending from an intersection in the grid (its initialization value) to some other point (its post-optimization value). As can be seen in the plot, the optimization process sometimes takes a seed far from its
initialization value, and sometimes it leaves it nearly unchanged. The circle at [Threshold=0.637, Gradience=1.61] represents the lowest MAD value across the entire plot - i.e. the 'correct answer'. Only two seeds reached this solution, neither of which had the #1 best fit among initialization values before optimization (instead being ranked #4 and #7). This fact underscores the importance of using multiple seeds to explore the search space.

While the parameter combination output from `optim()` usually has a better fit than the original seed one used to initialize the search, it occasionally does not. As such, the fit for both pre- and post-optimization for each of the 10 seeds are compared in order to determine which has the lowest weighted MAD (hence which transition shape is ultimately selected for use in the stylization). The following figure illustrates the final result with the same rising F0 transition represented in the above heat plot. Over the grey F0 points, the optimized contour-fit is superimposed in black.

![Figure 3.12: Example comparison of ultimately-selected transition shape with original F0 data](image)

Figure 3.12: Example comparison of ultimately-selected transition shape with original F0 data
Overall, the stylization fits the raw F0 data very tightly, suggesting a possible contour shape 'underlying' the various kinds of noise in the F0 track. Indeed, most of what is not retained by the stylized shape appears to be transient perturbations, e.g. the brief fall around 5500 ms after the [ð] of 'the'. Note also the brief F0 halving near the end of the transition. Such F0-tracking artifacts merely add a cluster of outlier points that affect all candidate transition shapes equally, and since the goodness-of-fit criterion is the \textit{median} absolute deviation, these generally do not compromise the results. The fact that the MAD is weighted by intensity can be seen in the fact that the high-intensity stretch at the end of the contour is captured quite closely at the sacrifice of the low-intensity stretch around 5400 ms.

3.5 **Storing and analyzing stylization data**

The following is the final stylization for the example utterance discussed throughout this section. Each row contains information about a vertex and the transition \textit{leading up to it}, and the numbers along the left side of the table index the transitions. Thus, the first row has an index of '0' and all information relating to the transitions is NA. The 'Index' column references which frame in the soundfile's F0 track a vertex corresponds to (e.g., '4369' means that the vertex in question corresponds to the 4369th frame in the F0 track). For ease of reference, Figure 3.5 above is repeated immediately below, with numbers corresponding to the transition indices from the table.
Table 3.4: Vertex and transition information stored in stylization for the example utterance

<table>
<thead>
<tr>
<th>Index</th>
<th>F0</th>
<th>Time (ms)</th>
<th>Threshold</th>
<th>Gradience</th>
<th>Weighted MAD</th>
<th>Voicing (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>4369</td>
<td>4381</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>1</td>
<td>4523</td>
<td>4535</td>
<td>0.096</td>
<td>0.08</td>
<td>0.208</td>
<td>100.0</td>
</tr>
<tr>
<td>2</td>
<td>4862</td>
<td>4874</td>
<td>0.860</td>
<td>1.69</td>
<td>0.742</td>
<td>80.6</td>
</tr>
<tr>
<td>3</td>
<td>4998</td>
<td>5010</td>
<td>0.463</td>
<td>0.64</td>
<td>0.302</td>
<td>100.0</td>
</tr>
<tr>
<td>4</td>
<td>5229</td>
<td>5241</td>
<td>0.584</td>
<td>2.74</td>
<td>0.214</td>
<td>48.7</td>
</tr>
<tr>
<td>5</td>
<td>5340</td>
<td>5352</td>
<td>0.588</td>
<td>0.88</td>
<td>0.411</td>
<td>99.1</td>
</tr>
<tr>
<td>6</td>
<td>5760</td>
<td>5772</td>
<td>0.637</td>
<td>1.61</td>
<td>0.214</td>
<td>72.2</td>
</tr>
<tr>
<td>7</td>
<td>5874</td>
<td>5886</td>
<td>0.880</td>
<td>0.06</td>
<td>0.161</td>
<td>100.0</td>
</tr>
</tbody>
</table>

Note how gradience indexes the extent to which a transition resembles a straight line. Thus, the first and last transition have very low gradience values (0.08 and 0.06 respectively) and are nearly perfectly straight, unlike transition #4 (with a relatively high gradience value of 2.47). The effects of threshold can be seen most clearly where gradience is high. The threshold value for transition #4 (0.584) is close to 0.5, hence a clear curvature can be observed at both ends of
the transition (i.e. near the beginning and the end). In contrast, the threshold value for transition #2 is relatively late (0.860), hence the first half has a gradual curvature and the second half quickly darts upward. The next column in Table 3.4 contains the weighted MAD for each transition, which can be used as an evaluation metric for the quality of the stylization. For example, if the weighted MAD surpasses a pre-determined cutoff (e.g. 1.00), the transition can be red-flagged as potentially problematic.

One final piece of information is also retained about each transition (shown in the last column of the table) - the percentage of frames within the time range of the transition that are voiced (i.e. have an F0 value). This information is used to determine the darkness and thickness of the lines. Transitions where most or all frames are voiced (such as #1, #3, and #7 in the above figure) are represented with a thick, dark black line. In contrast, transitions with many voiceless frames (such as #4 above, with over half voiceless frames) are represented with a thin, faint grey line. In cases where the speaker pauses in the middle of their utterance (hence creates a region of mostly almost entirely voiceless frames), this kind of the data in the stylization can be used to filter out such 'false' transitions by setting a threshold criterion (e.g. treat everything under 25% as a pause). This information is also captured in the rich visualization as well, which makes it clear at a glance which transitions have more voiceless frames.

Since stylizations can be stored in matrix form (as in Table 3.4), they can be easily exchanged among researchers (e.g. as tab-delimited textfiles). This is analogous to how how ToBI annotations can be exchanged among researchers in Praat TextGrid format (or, earlier, as TextTier and IntervalTier annotations in ESPS xwaves). Making stylizations 'portable' in this way helps ensure one's analyses are reproducible by others researchers.
Once a set of soundfiles have been stylized in this way, they can be 'queried' to explore research questions of interest. For example, it may be asked, "What percent of the time does a high target (i.e. the endpoint of a rising F0 transition) occur within ___ (some specified target word/syllable)?" Since the stylization retains the rich phonetic details of the F0 contour, it is also easy to investigate follow-up questions such as "What sort of distribution is observed in the scaling/alignment of that high target?" or "What is the shape of the transition preceding/following that high target?" Such queries draw upon the quantitative representation of the F0 contour stored in the stylization object, and as such are formally explicit in nature. As such, queries can also be easily exchanged from one researcher to the next, either by sharing code directly or by using pseudo-code to summarize the gist of the query in a publication. For an illustration of these ideas, see Chapter 6 for the set of queries used in the present dissertation.

With the typology (Chapter 2) and methods (Chapter 3) thus established, the stage is set for delving into the third and final piece of the present dissertation: the empirical component. The following chapter discusses the specific intonational phenomena of interest, thereby generating three research questions to be tested in the remainder of the dissertation.
Chapter 4: Research questions

This chapter is structured as follows. First, a broad overview of the intonation systems of Japanese and English is provided in section 4.1. Then, in section 4.2, the two systems are contrasted in greater detail, generating three research questions based on three structural differences between the two languages (hence three areas where L1 transfer may be predicted to occur in learners' L2 production).

4.1 Description of the intonation system of Japanese and English

The following two sections sketch out the basic facts of how Tokyo Japanese intonation (§4.1.1) and American English intonation (§4.1.2) is modeled in Autosegmental-Metrical (AM) terms. Each section first focuses around a discussion of a schematic diagram summarizing the overall intonational grammar for the relevant language. This is then followed by a walkthrough of several utterances exemplifying the various components to this diagram. The utterances used for this purpose are taken from the official set of 'canonical' soundfiles for each language, distributed in order to train transcribers how to annotate using the relevant ToBI system. These soundfiles will be visualized using the rich visualization approach documented in §3.1, but instead of stylizations, the officially-distributed transcriptions for each soundfile will be used instead.
It is worth noting that the details of the intonation systems of these languages have been the subject of much debate. (See, for example, the proposed revisions to Japanese and English in Gussenhoven (2004) and the fundamental re-analysis in Dilley (2005).) However, since this section sets out only to present the basic facts of each system, a comprehensive synthesis of such disagreements is beyond the present purposes. As such, the discussion in this section is limited to the phonological model underlying the latest version of the transcription system for each language.

4.1.1 Broad overview of the AM model of Japanese intonation

The most recent instantiation of the AM model of Tokyo Japanese intonation is captured in the X-JToBI framework (Maekawa et al. 2002). This system is a revised and expanded version of the earlier J_ToBI framework by Venditti (1997), whose conventions were largely based on Pierrehumbert & Beckman (1988). (See Venditti (2005) for an overview of the history of the framework.) Compared to J_ToBI, the main new characteristics of X-JToBI are "1) Exact match between the time-stamp of tone labels and the timing of physical events, 2) Enlargement of the inventory of boundary pitch movements, 3) Extension and ramification of the usage of break indices, and 4) Newly defined labels for filled-pause and non-lexical prominence" (Maekawa et al. 2002). However, none of these changes are crucial in the present context. The only thing of consequence for the present discussion is the core elements of the system, which were essentially unchanged between the two versions of the framework. Thus, the discussion below follows the original, simpler J_ToBI system (Venditti 1997), whose accompanying soundfiles and annotations have been used to create the figures for the examples.
Figure 4.1 below is a 'state diagram' (a kind of directed graph often used in describing finite state machines) that summarizes the possible sequences of tones Tokyo Japanese as captured by J_ToBI. The diagram is a much expanded version of the one for J_ToBI presented in Noguchi et al. (1999) and resembles the one for X-JToBI in Igarashi et al. (2013). Note that since the state diagram only shows the relationship between the labels on the tone tier, other kinds of information contained in the annotations (such as break indices) are not represented.

![State Diagram](image)

**Figure 4.1: State diagram representing possible sequences of tones in J_ToBI**

Any path through the diagram (representing a single utterance) must start at the leftmost node (%L). A path that proceeds directly to the rightmost node would represent an utterance consisting of a single minor (or 'accentual') phrase - the level prosodic structuring immediately above the phonological word (Pierrehumbert & Beckman 1988:117-118). The diagram's leftward-pointing arrow (from L% to H-) creates recursivity, thus allowing for utterances with multiple minor phrases (each ending with a L%). A series of multiple minor phrases over which downstep (iterative narrowing of the pitch range) occurs at each pitch accent is referred to as a major (or 'intermediate') phrase. For terminological convenience, 'minor phrase' will be

---

1 This diagram has been algorithmically verified to capture every label on the tone tier in every J_ToBI canonical file (both example and practice utterances) as well as every pairwise transition between them (i.e. in a bigram model of the tone labels). The only exceptions are the following, each of which has been omitted: (1) the '*' label (indicating the transcriber is uncertain whether an accent is present), (2) the '>' and '<' symbols (marking early/late alignment of an F0 feature), (3) labels for complex boundary pitch movements (e.g. L%HL%), (4) %wL and wL% ('weak' low boundary tones, treated here as simply %L and L%), (5) the transition L% → %L (spanning two separate 'utterances'), and (6) transitions that are due to disfluencies or typos.
temporarily abbreviated to simply 'phrase' throughout the rest of the discussion in this section. However, the distinction between major and minor phrases becomes important later below (in §4.2.2).

Each node in the state diagram will now be discussed in turn.

1. **Initial %L:** In principle, every utterance begins with an initial low target. The vertical scaling (or 'strength') of this target varies depending on factors such as the location of the accent in the phrase and the initial syllable's weight.

2. **H-:** The %L is then generally followed by a high tone (H-).² Where this tone is phonologically associated depends on the weight of the initial syllable in the phrase. If it is heavy, the H- aligns with the first sonorant mora (at least in fast speech and/or utterance-medial position), whereas if it is light the H- aligns with the second sonorant mora (Pierrehumbert & Beckman 1988:127). In either case, the anchoring point for the H- is defined with respect to a phrase boundary, thus signaling the beginning of each new phrase. Accordingly, this tone is referred to as 'phrasal H-'. This tone is marked with a '-' rather than a '%' (i.e. H- rather than %H) to indicate that it is associated to a position slightly removed from the phrase boundary itself (here, 1 or 2 moras away).

3. **H*+L:** The next tone depends on whether the phrase contains an accent (whose presence-vs.-absence and location is determined through a combination of lexical specification and

---

² In phrases where the H*+L accent occurs on the first mora, the phrasal H- does not surface. More precisely, in such cases, the high target for the H- is absent, and instead there is only the fall for the H*+L accent. Pierrehumbert & Beckman (1988:126-134) describe how this phenomenon can be accounted for as a process whereby the linking between the segmental string and the phrasal H- is phonologically 'blocked'. This results in %L transitioning directly H*+L - a possibility not included in Figure 4.1. Note also that even if a H*+L accent occurs on the second or third mora, the H- may not be visibly distinguishable from it in an F0 track. In such cases, Venditti (1997:4-5) prescribes not to mark the phrasal H- in an intonation transcription.
numerous phonological processes). If the phrase does contain an accent (i.e., the phrase is 'accented'), then a high target will occur near that syllable followed shortly thereafter by a low target, forming the $H^*+L$ pitch accent (the only possible phonological pitch accent shape in Japanese). If the phrase does not contain an accent (i.e., the phrase is 'unaccented'), then the peak F0 across the phrase occurs at the phrasal H-.

4. **Phrasal L%**: Next is a low target that is obligatory at the end of every phrase (hence the name 'phrasal L%'), whose scaling varies in the same way as described above for Initial %L. In an accented phrase, the F0 is already low from the low target of the $H^*+L$, hence it gently falls slightly further down until the end of the phrase. If the phrase has no pitch accent (i.e., it is 'unaccented'), the phrasal H- forms a peak, after which the F0 slopes down to the phrase's end. In a longer utterances with multiple phrases, each phrase follows the sequence \{ $H^* \rightarrow (H^*+L) \rightarrow L%$ \}.

5. **H%/Ø**: A chain of one or more such phrases can terminate with a high target (H%), thus creating an F0 rise from the preceding L%. The presence of the boundary H% prevents final lowering from occurring, which can be thought of as a 'final raising' effect (Pierrehumbert & Beckman 1988:72-76). Since these two tones are associated with the same prosodic boundary (at the right edge of the last phrase), this rise typically begins from within the final syllable. If there is no H% (hence this location is phonologically null, cf. the Ø in the last node of the diagram), the chain of phrases simply ends with the low target of the L% from the last phrase.

To illustrate the various arrows in the state diagram, three utterances from the J_ToBI canonical files will now be discussed. The basic structure of single phrase is illustrated in the
following figure. The utterance depicted therein represents the beginning of the J_ToBI canonical file *nibanme1*, where it occurs in the context in (1). (Tone labels are only shown for the first phrase; the second phrase is included only for phonetic context.)

![Spectrogram and waveform of utterance *nibanme1*](image)

Figure 4.2: Excerpt of J_ToBI example utterance 'nibanme1'

(1)  nibanmé-no   shinshitsu-no   (mádo-wa...)  
second-GENITIVE  bedroom-GENITIVE  (window-TOPIC...)  
(as for) the second bedroom (window...)

The first phrase begins with the obligatory initial low target (%L). This then rises to a high target on the second mora of the phrase, i.e. the [ba] in *ni-ba-n-mé-no*. In the spectrogram, this target occurs near the middle of the vocalic region for the [a]. Since the word *nibanmé* is accented on

---

3 In this and all following rich visualization plots in this chapter, the tone labels are taken directly from the official annotation of each file, as are the timestamps for the word boundaries and tone labels. (The Romanization scheme has been adjusted to coincide with the system used throughout the present dissertation.) Note, however, that various other aspects of those annotations that are not relevant in the present context - most notably break indices - are omitted from the plots.

98
mé, the high target for the H*+L accent occurs in the syllable [me]. Here again, this occurs near the middle of the vocalic region for the [e]. This is followed by a fall to a low target that simultaneously represents two different phonological tones: the trailing low target component of the accent (i.e. the '+L' of H*+L) and the L% tone marking the end of the phrase. These two tones can be thought of as phonetically 'fused', with the low target representing the 'cumulative exponence' (to borrow a term from morphology) of the two tones. These two tones are 'fused' into one low target in this case simply because there is not enough segmental material separating them. When the accent occurs several syllables away from the end of a phrase, two distinct low targets can be distinguished (Pierrehumbert & Beckman 1988). Finally, the first phrase ends with a 'boundary rise' to a high target (H%), marking the division between this phrase and the next one within the larger utterance. In sum, this utterance illustrates the sequence of all five 'nodes' in the diagram in Figure 4.1, left to right, in order.

Whereas the above figure illustrates the basic structure of a single phrase, the following figure illustrates a sequence of multiple phrases (cf. the curved leftward arrow at the bottom of Figure 4.1). The utterance depicted therein represents the first two phrases in the second half of the J_ToBI canonical file hachiue, whose meaning is provided in (2).4

---

4 The full sentence is as follows, making it clear why the file is called hachiue '(potted) plant':

hachiue-wa ē ki-to ēto shōmén-no iriguchi-no aida-ni kuru-yōni okimasu
plant-TOPIC um tree-and um front.side-GEN entrance-GEN between-in come-so.that put
I'll put the plant, um, so that it comes in between the tree and, um, the front entrance.
Figure 4.3: Excerpt of J_Tobi practice utterance 'hachue'

(2) ēto shōmén-no iriguchi-no...
    um front.side-GENITIVE entrance-GENITIVE
    um, the front entrance's ...

The structure of the first phrase, shōmén-no, is similar to the one just discussed, consisting of an
initial low target, a phrasal H-, a H*+L accent, and a L% boundary tone.5 The initial low target is
realized at the end of the preceding filled pause ēto 'um', the phrasal H- inside the [o] of the
heavy (bimoraic) syllable shō, the H*+L inside the [e] of the lexically accented syllable mén, and
the L% at the transition from [o] to [i] at the word boundary. The second phrase has the same
structure as the first except that it lacks the H*+L since the phrase is unaccented. (This represents

5 The use of the label '%wL' rather than '%L' either indicates that the first syllable in the phrase is heavy (as here) or
bears a H*+L accent. Both of those factors cause the low target to be realized with a higher F0 (i.e. a 'weaker' low
value), hence the term '%wL' (weak L%).
the curved rightward arrow 'skipping' the H*+L in Figure 4.1.) This configuration creates a L% H- L% sequence, where the H- is realized in the syllable ri.

### 4.1.2 Broad overview of the AM model of English intonation

The description of English intonation in the section below follows the latest version of the labeling guidelines for Mainstream American English (MAE) ToBI (Beckman & Ayers-Elam 1997) and its subsequent re-packaging in a more pedagogical format (Veilleux et al. 2006). (See Beckman et al. (2005) for an overview of the history of the framework.) The soundfiles and annotations have been used to create the figures for the examples come from the official distribution by Gravano (2010).

In the same way as Figure 4.1 above, the following figure summarizes the possible sequences of tones in MAE_ToBI as a state diagram. The diagram is a revised version of the one originally presented in Pierrehumbert (1980).

![State diagram representing possible sequences of tones in MAE_ToBI](image)

---

6 As before, this diagram has been algorithmically verified to capture every label on the tone tier in every MAE_ToBI canonical file as well as every pairwise transition between them. The only exceptions are the following, each of which has been omitted: (1) the downstepped tones !H*, L+!H*, L*+!H, and !H- (treated here as their non-downstepped equivalents), (2) the four labels indicating transcriber uncertainty (*?, X*?, X-?, X%?), (3) the '<' symbol (marking late alignment of an F0 feature), (4) the 'HiF0' label (marking the actual F0 maximum irrespective of segmental location), (5) the '%r' label (marking a reset of the F0 range), and (6) transitions spanning two separate 'utterances' (e.g. L% → %H).
Each node in the state diagram will now be discussed in turn. Throughout the discussion, the term 'phrase' will refer to the domain of one or more pitch accents plus a phrase tone (H- or L-). Within a given phrase, downstepping (iterative narrowing of the pitch range) frequently occurs from one pitch accent to the next. A given utterance can have multiple phrases by virtue of the leftward-pointing arrow at the bottom of the diagram.

1. **%H/Ø**: In pragmatically marked circumstances (e.g. to communicate a contradiction to what the interlocutor has just said), utterances can begin with a high target (%H). Normally, however, this location is phonologically null (Ø), and the F0 for the utterance simply begins near the middle of the speaker's habitual pitch range.

2. **L*/H*/L+H*/L*+H/H+!H***: Next is a string of one or more pitch accents, since multiple pitch accents can occur next to each other within the same phrase. Each of the pitch accents can be one of many different possible shapes. The five in Figure 4.4 are those included in latest version of MAE_ToBI. As mentioned in Chapter 1 (§1.1.2), the last pitch accent in a phrase is referred to as the 'nuclear' pitch accent (or 'nucleus') and plays a special role in the information structure of many languages.

3. **H-/L-**: Following this sequence of pitch accents is either a high or low target for the phrase tone.\(^7\) In longer utterances with multiple phrases, each individual phrase ends with a phrase tone, hence multiple phrase tones can be spread across a single utterance. While phrase tones are phonologically associated with the end of the phrase (hence their name), they are "not necessarily localized at the phrase edge. Rather, when the nuclear accent is far from the end

---

\(^7\) While the term 'phrase accent' is the more standard label for this kind of tone, to maximize the terminological distinction from the unrelated construct of 'pitch accent', the term 'phrase tone' is used instead throughout the present dissertation.
of the intermediate phrase, the phrase [tone] fills in the space in between it and the phrase edge, creating a long flat valley for L- realized over a long stretch [...] or a long plateau-like region for H- realized over a long stretch" (Beckman & Ayers-Elam 1997:18). Since a phrase tone is removed from the phrase boundary itself, it is marked with a '-' (rather than a '%'), like the case of Japanese H- discussed above.

4. H%/L%: Finally, the utterance terminates with a high or low boundary tone. The traditional account (dating back to Pierrehumbert 1980) is that a H- phrase tone 'upsteps' the H%, thus making the tone labels in this position somewhat non-transparent. For example, following a H-, a L% is raised to the same height as the H-, thus creating a flat plateau. Likewise, following a H-, a H% is raised even higher, hence moderate-sized rises are transcribed as L-H% and large rises are transcribed as H-H%. Thus, the H% vs. L% distinction is essentially just a privative indicator of whether the final portion of the contour involves a rise, and the preceding phrase tone (H- or L-) creates the register at which this contour is realized.

To illustrate the various arrows in the state diagram, three utterances from the MAE_ToBI canonical files will now be discussed. The first utterance illustrates the structure of a single-phrase utterance.
The utterance begins with a %H on the unstressed syllable 'do', thus creating the pragmatically marked 'contradiction contour'. The F0 then falls to the low target (L*) in the first vowel (before the [l]) in 'really'. In this utterance, the low target for the L- phrase tone aligns with the utterance-final syllable, thus creating a low valley from the L* to the L-. The F0 then rises to the final high target for the H% at the end of the sonorous portion of 'one'. The tones making up this utterance represent a relatively simple path through the state diagram in Figure 4.4, progressing left to right with one tone for each of the four nodes.

The second utterance illustrates one kind of recursivity inherent in the state diagram - namely, how multiple pitch accents can occur in sequence inside a single intonational phrase. It represents the first intonational phrase from the longer utterance, *State law now requires public construction projects to set aside one percent of their budgets for artwork.*
Since this utterance has no initial boundary tone, the first target is for the H* on 'state'. This is then followed by a downstepped high target (!H*) on 'law', followed by another high target (!H*) that is downstepped even further on 'now'. (The traditional account for the intervening dips is that they represent 'sagging transitions' between high targets.) Next, the F0 drops even further to a low target at the end of [ɹә] (in 'requires') followed by a high target on [κwai]. These two targets collectively represent a L+H* pitch accent associated with the syllable [κwai]. Finally, the utterance ends with a fall to a low target, representing the cumulative exponence of both the L- phrase tone and the L% boundary tone. What is crucial in the present context is the string H* !H* !H* L+H* in this utterance, illustrating how it is possible to have multiple neighboring pitch accents, one after another, in Mainstream American English. This therefore illustrates the leftward arrow at the top-center of Figure 4.4, whereby one pitch accent is followed by another.
The third and final utterance illustrates the other kind of recursivity in the state diagram, whereby a single utterance can contain several smaller phrases, each ending with a phrase tone.

Figure 4.7: The MAE_ToBI canonical file 'lazy'

This utterance consists of three phrases: { He's lazy } { and crazy } { and stupid }. The first and second phrases contain low targets (L*), aligned to the stressed first syllable of 'lazy' and 'crazy', respectively. These then rise to a high phrase tone at the relevant phrase boundaries (i.e. the ending points of the same two words - 'lazy' and 'crazy'). Such sequences of multiple instances of L* H- are what is usually referred to as 'listing intonation' (here imparting the connotation that the speaker could have continued listing multiple additional personality flaws of the person in question). The F0 then resets to the middle of the speaker's pitch range in the word 'and' (i.e. the beginning of the final phrase). After reaching the high target on the stressed syllable [tu] of 'stupid', the F0 falls to the end of the utterance (L-L%). (Amid the many black pixels (i.e. secondary candidates) due to the creaky voice, a few light-grey points (i.e. low-strength primary
candidates) can be distinguished in the vowel of the final syllable [pid] near 150 Hz.) The notable feature of this utterance is that it has a total of three different phrase tones (H-, H-, and L-), each marking the end of a different phrase. Thus, this utterance exemplifies the leftward arrow at the bottom of Figure 4.4, whereby a phrase tone is followed by a pitch accent.

4.2 Research questions

The survey of the intonation systems of English and Japanese just presented makes it clear that there are many similarities between the two systems (Beckman & Pierrehumbert 1986). However, there are also several key structural differences between the two systems, three of which are discussed in the following sections (§4.2.1 through §4.2.3). Each section presents a contrastive analyses of the relevant aspect of the two languages’ intonation systems.

These three cross-linguistic differences form the basis of the three research questions explored in the present study. Each research question explores the implications of one of these differences for L2 acquisition, examining some way that Japanese EFL learners may be expected to "speak English with Japanese intonation". Since the three chosen differences represent disparate types of mismatches in terms of the typology sketched out in Chapter 2, the three research questions collectively represent a wide range of different classes of L2 intonation transfer.

All three research questions share a common structure: "how often do Japanese EFL learners do ___?" This makes it possible to assemble the results for all three research questions in terms of a broader hierarchy (i.e. which kinds of transfer occur more/less often). This higher-level analysis is presented in Chapter 7 (§7.1) Of course, as addressed in §1.1.4, transfer is often not the only explanation for a learner's divergence from native norms. In this light, each research
question below simply asks whether the phenomenon in question occurs systematically in the contexts that would be predicted based on transfer. In doing so, the goal is merely to demonstrate the validity of transfer as a possible explanation for the observed phenomena.\(^8\)

### 4.2.1 Research question 1

Recall from Section 4.1.1 that the left edge of each prosodic ('accentual') phrase in Japanese is demarcated with a phrasal H-. Three examples of this tone were seen above, illustrating the different sites where this tone can be aligned. In Figure 4.2, the H- occurs on the second mora (ba) in an accented phrase (nibanmé-no) beginning with a light syllable (ni). In the first phrase in Figure 4.3, the H- occurred on the first mora (sho) in an accented phrase (shōmèn-no) beginning with a heavy syllable (shō). In the second phrase in Figure 4.3, the H- occurred on the second mora (ri) in an unaccented phrase (iriguchi-no), where it formed the F0 peak of the phrase as a whole. In all three cases, the H- creates an independently identifiable turning point at the left edge of the prosodic phrase, thus marking the boundary between adjacent phrases. The presence of this tone is effectively guaranteed by the intonational grammar of Tokyo Japanese in Figure 4.1, where it obligatorily occurs at the beginning of each phrase (before the H*+L pitch accent, if any).

In contrast, English has no tone analogous to the phrasal H-. The English state diagram in Figure 4.4 makes it clear that this position is phonologically empty in English. Note that there

---

\(^8\) From the perspective of Jarvis (2000), what is missing from the present study's design is empirical data from learners with other L1s that (1) differ from Japanese in the relevant ways and (2) demonstrate that they do not show the same patterns as Japanese EFL learners. Such data would put claims of transfer on much more solid ground. It is left to future research to complete the picture with this cornerstone of the argument.
was no such high target at the beginning (before the first pitch accent) of any of the utterances discussed above in section 4.1.2 (Figures 4.5, 4.6, and 4.7).

The closest analogue English has to Japanese H- is %H (mentioned in 4.1.2 and exemplified in Figure 4.5), since both are high tones associated with the left edge of a phrase. However, these two differ in at least four ways. First, English %H occurs at the beginning of an entire utterance, whereas Japanese H- occurs at the beginning of almost every phrase. Second, English %H is always associated with the absolute beginning of the utterance, whereas Japanese H- is usually one mora removed from the phrase boundary (except for the case mentioned in the second footnote in this chapter). Third, English %H is usually realized as a fall from the top of the speaker's pitch range onto a low target, whereas Japanese H- is usually realized as an elbow during a rise onto a high target. Fourth, English phrasal %H is rare and pragmatically marked (e.g. indicating contradiction), whereas Japanese H- is frequent and pragmatically unmarked, resulting from an automatic 'post-lexical' phonological process (free of pragmatic function and lexical properties). Thus, despite their superficial similarity, these two are indeed distinct, thus reinforcing the premise that English indeed has nothing comparable to Japanese H-.

The fact that the left edges of prosodic phrases are generally not marked intonationally in English is illustrated in the following figure, where the first pitch accent in the phrase comes several words after the beginning of the utterance.9

---

9 The label 'HiF0' (marking the location of the locally highest F0) at 665 ms in the official ToBI TextGrid for this file has been omitted from this figure.
Crucially, in this figure, over the region before the first pitch accent, the F0 merely interpolates from the middle of the speaker's pitch range towards the high target for the H*. This contrasts with Figure 4.2, where there is a readily identifiable elbow in the F0 track corresponding to the phrasal H-.

While the utterance in Figure 4.8 is similar to Japanese in that the first target in the pitch accent is high (H* in English and H*+L in Japanese), the same pattern holds true even if the target for the pitch accent is low. This is exemplified in the following figure, representing the first half of the larger utterance "Will you have marmalade? Or jam?":

*Figure 4.8: First utterance inside the MAE_ToBI canonical file 'tags1'*

The graph shows the F0 track over time, with pitch accents labeled as H+, L, and L-H%. The corresponding text is shown below the graph, highlighting the pitch accent labels and the time frame. The F0 values range from 0 to 4500 Hz, and the time is marked in milliseconds from 400 to 1200 ms.
Here again, the F0 interpolates from the middle of the speaker's pitch range - this time downward towards the low target for the L*. Crucially, note the lack of tonal specification before the L* - something would be impossible if there were something like a phrasal H- in this utterance.

To summarize, Japanese phrases generally begin with a post-lexical phrasal H- before the H*+L pitch accent (if any), whereas in English phrases there is generally no tone before the first pitch accent. From the perspective of a Japanese EFL learner, this means that an L1 category (phrasal H-) has no close neighbors in the L2 category set. This effectively creates a 'gap' in a region of the contour where, in the L1, there is normally an F0 target. In terms of the typology from Chapter 2, this constitutes a [Position, Emptiness] mismatch, where the position 'initial phrase tone' is occupied in the L1 (Japanese) but not in the L2 (English). Given the 'automatic' post-lexical status of this L1 phrasal H- category, it is possible that a Japanese EFL learner
would use an L1-transferred phrasal H- before the first pitch accent in their L2 English utterances. Whether this is indeed the case is the topic of the first research question:

(3) **Research Question 1**: In their L2 English production, how often do Japanese EFL learners exhibit evidence for an L1-transferred phrasal H-?

The exact operationalization of 'evidence for an L1-transferred phrasal H-' is presented in detail in Chapter 6 (§6.1.1). In broad terms, such evidence takes the form of a high F0 target at the left edge of a prosodic phrase (before its first pitch accent). Based on the facts sketched out earlier above in this section, it may be hypothesized that F0 targets matching this description should occur for many Japanese EFL learners but not for native English speakers.

**4.2.2 Research question 2**

The next research question concerns the difference in looping structure between the English and Japanese state diagrams. Recall from Section 4.1.1 that, in longer Japanese utterances with multiple minor phrases, each follows the sequence \{ H- → (H*+L) → L% \}. Inside this looping domain, it is structurally guaranteed by the intonational grammar that there can only be a maximum of one pitch accent (either exactly one or none at all). In other words, it is disallowed to have more than one H*+L pitch accent occur directly next to each other inside the same minor phrase. This distributional restriction is illustrated in the following figure:
Figure 4.10: First half of the J_ToBI canonical file 'sankaku'

(4) sánkaku-no yáne-no (mannaka-ni okimásu)
triangle-GENITIVE roof-GENITIVE center-in place
(I will place it right in the center) of the triangle roof.

This utterance consists of two minor phrases: \{sánkaku-no\} and \{yáne-no\}. If they were phrased together as \{sánkaku-no yáne-no\}, there would be no (w)L% at around 800 ms to keep the F0 low until the end of the (sankaku)-no, and the F0 would begin rising during sankaku as soon as the drop for the H*+L was completed. The fact that this does not occur is evidence for the existence of the (w)L%, hence the fact this utterance is divided into two minor phrases.

The case of English is somewhat different from that of Japanese. English lacks a similar restriction on the number of pitch accents per phrase, hence it is perfectly normal (and indeed quite frequent) for a prosodic phrase in English to contain multiple pitch accents. In terms of the state diagram in Figure 4.4, this is possible because the looping domain for the prosodic phrase
(leftward pointing arrow at bottom of figure) encompasses a smaller loop (leftward pointing arrow at top of figure) allowing for recursivity in the string of adjacent pitch accents. A phrase of such as large size would correspond to a Japanese major phrase, since both can contain multiple pitch accents. Furthermore, recall that Figure 4.6 ("state law now requires...") had as many as four pitch accents (H*, !H*, !H*, L+H*) inside the same phrase, each downstepping the other. This is also akin to the Japanese major phrase, since in both languages a phrase of this size serves as the domain for downstep.

It is important to note that this discrepancy between the two languages cannot be attributed to a difference in inventory alone. Both languages have similar tones - L% in Japanese and L- in English - both of which serve to demarcate the end of prosodic phrases. The difference is the size of phrase they attach to: minor phrases in Japanese and (the Japanese equivalent of) major phrases in English.

Another way of thinking about one language having many pitch accents per major phrase is that it lacks a tonal marker between minor phrases. English falls under this category; hence, the language can be treated as lacking tonal marking at the level equivalent to the Japanese minor phrase. (Note how no edge tone is required after each pitch accent in Figure 4.4.) As a secondary consequence of this state of affairs, an equally-sized utterance in the two languages would have systematically more (low) edge tones in Japanese than in English. As such, in terms of the typology in Chapter 2, this cross-linguistic difference can be classified as a [Density, Edge tones] mismatch.

As was seen above for the first research question, this difference could also conceivably have implications for L2 acquisition. The difference predicts that Japanese EFL learners (with a
high-density L1 and low-density L2) may frequently parse a long utterance into many minor phrases (rather than a single major phrase with multiple pitch accents). In other words, since a L% must intervene to prevent multiple pitch accents from being next to each other in the L1, learners with this L1 background might 'chop up' a larger utterances into many smaller minor phrases with L% tones. This prediction is tested with the following research question:

(5) **Research Question 2**: In their L2 English production, how often do Japanese EFL learners use low F0 targets at prosodic boundaries?

As mentioned above, since both languages have analogous tones, the prediction is not that learners will be the only ones to show this pattern (to the exclusion of native speakers). Rather, what is crucial is the frequency with which these low boundary tones are used, such that learners are hypothesized to do so disproportionately often. In a sense, the expected pattern is a form of tonal epenthesis. Just like a learner whose L1 only has simplex onsets may epenthesize many additional vocalic portions when producing an L2 that has complex onsets, Japanese EFL learners may be expected to epenthesize many additional low F0 targets at phrase boundaries.

### 4.2.3 Research question 3

A final major difference between English and Japanese lies in the temporal alignment of the point at which F0 begins to rise at a prosodic boundary (e.g. to indicate a question, among numerous other possible functions). Recall from the discussion above in §4.1.1 that every phrase in Japanese obligatorily ends with a low F0 target representing the 'phrasal L%' - a boundary tone that is phonologically associated with the right edge of the phrase. If an additional high F0 target (e.g. H%) is added to form a boundary rise, then the low target aligns as early as possible within the final syllable and the rise occupies the rest of that syllable. In a sense, the phonological association of the phrasal L% with the phrase boundary effectively guarantees this
fixed alignment of the beginning of the boundary rise. This holds true regardless of the function of the boundary rise (e.g. question, incredulity, insisting, etc.), utterance position (medial vs. final), whether the phrase is accented, and (in accented phrases) which syllable bears the H*+L accent. (Note how, in the state diagram in Figure 4.1, a H% must always be preceded by a L%.)

This set of facts is illustrated in Figures 4.11 and 4.12 below. First, Figure 4.11 (representing the canonical J_ToBI utterance in (6)) illustrates the case of a boundary rise on an accented phrase. In this and the following figure, the ‘<’ is the J_ToBI symbol used to mark the location of the actual F0 peak associated with the preceding high tone (here, H-). The symbol ‘>’, on the other hand, indicates the highest F0 point (seen by the original transcriber) in a boundary pitch movement such as the boundary rises here.

![Figure 4.11: The J_ToBI canonical file 'mayumi'](image_url)

*Figure 4.11: The J_ToBI canonical file 'mayumi'*
(6) Mayumi mo nósnda yo
Mayumi also drank ASSERTION
Mayumi drank, too.

In the figure, the boundary rise begins near 700 ms, at the start of the final syllable \textit{yo}. The turning point at that location in the F0 contour (here, a local minimum) is modeled as a L\% boundary tone. Note that this point occurs multiple syllables after the accented syllable \textit{nón}.

Figure 4.12 (representing the canonical J\_ToBI utterance in (7)) illustrates the case of a boundary rise on an unaccented phrase.\textsuperscript{10}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure412.png}
\caption{The J\_ToBI canonical file 'kazumi'\textsuperscript{10}}
\end{figure}

(7) Kazumi ga yonda yo
Kazumi NOMINATIVE called ASSERTION
Kazumi called.

\textsuperscript{10} Note that, due to a phonological neutralization process, the pattern in Figure 4.12 can also occur on an accented phrase if the accent falls on a phrase-final syllable (e.g. in an interrogative rendition of \textit{otokó} 'man?').
Here again, the boundary rise begins at the start of the final syllable yo near 700 ms. The elbow at the end of the flat (if not slightly decreasing) stretch is modeled as a L% boundary tone. Since the F0 does not fall back down to its level from the beginning of the utterance (approximately 175 Hz), the L% can be thought as 'upstepped' (i.e. having undergone an upward register shift), akin to the traditional analysis of English L% as being upstepped after H- (Pierrehumbert & Beckman 1988:73-74). Since this contour contains two rises (one from %L to H- and another from L% to H%), this will be referred to as a 'double rise contour' in the present dissertation.11

While all examples of boundary rises discussed thus far occurred over monosyllabic particles (-no 'GENITIVE') in Figure 4.2 and -yo 'ASSERTION' in Figures 4.11 and 4.12), this is not a requirement. Even when produced over the bare nouns and verbs without the corresponding particles (nibanne 'second', nonda 'drank', and yonda 'called', respectively), the boundary rises would still begin from the final syllable. Thus, the relevant factor is not the morphological makeup of the segmental 'text' but rather its prosodic-phonological structure (namely, the location of the phrase-final syllable).

Boundary rises in English are much more complicated, as various combinations of phonological categories can yield rises of different shapes. (Note the numerous possible combinations of pitch accent and phrase tone that could precede a H% in the state diagram in Figure 4.4.) Moreover, if there are too few post-nuclear syllables to realize the tune over, this set of contrasts can be partially neutralized (e.g. L* L- H% vs. L* H- H% with 0 or 1 post-nuclear

11 Another analysis (e.g. Gussenhoven (2004)) would maintain that the L% is missing altogether in such double-rise contours. Instead, the flat portion could perhaps be analyzed as constituting a rightward tonal spread from the phrasal H- until the final syllable. This only constitutes a difference in underlying phonological analysis, however; both accounts agree on the basic phonetic alignment facts. As such, for tradition's sake, the discussion throughout the present dissertation will follow the original Pierrehumbert & Beckman (1988) account.
syllables). Of the larger system, what is crucial in the present context is that boundary rises in English frequently begin from the nuclear pitch accent, especially when there are few post-nuclear syllables (from 0 to about 3). In such cases, the boundary rise can potentially begin from a point earlier than the phrase-final syllable. One such case is illustrated in the following figure.

In this phrase, the nuclear stressed syllable [maɹ] is around 4650 to 4850 ms. As indicated by the location of the second 'L*' (just before 4800 ms), the boundary rise begins from inside this syllable (i.e. before the weakening in the spectrogram for the [m] at 4850 ms).12

![Figure 4.13: Third utterance inside the MAE_ToBI canonical file 'made3'](image)

---

12 Note that the final part of this contour is transcribed L* H- H%, where the phrase tone (H-) and boundary tone (H%) are jointly realized as a single high target. Depending on several factors, boundary rises from a nuclear stressed syllable over a small number of post-nuclear syllables can be transcribed in various other ways, such as L* L- H%, H* H- H%, or (if utterance-medial) L* H-. Despite the different transcriptions, it is an open question how many distinct phonological categories these represent. See Levis (2002) and Dilley (2007) for relevant discussion.
Thus, if the nuclear accent occurs one or more syllables from the right edge of the phrase, then the boundary rise can span multiple syllables. Of course, this does not mean that boundary rises always must do so. Various sequences of categories can create rises spanning only the phrase-final syllable, e.g. if the nuclear accent is on the phrase-final syllable. For example, if the utterance in Figure 4.9 above used the word 'jam' instead of 'marmalade' (hence "Marianna made the jam?"), the rise would occur over the single syllable "jam".

A double-rise pattern similar to the one observed for Japanese in Figure 4.12 is also possible in English. In such utterances, the final rise generally begins from the final syllable, much like in Japanese. These facts are exemplified by the final utterance.

Figure 4.14: MAE ToBI canonical utterance 'manitowoc'

Note that the relevant part of the transcription is identical between this utterance and Figure 4.13 above. The crucial difference is the number of post-nuclear syllables in the 'tail' of
the contour: two (*ma-lade*) in Figure 4.13 and nine (*ni-to-woc have a bow-ling a-lley*) here. The abundant segmental material between the nuclear pitch accent and the endpoint of the utterance creates enough space that two distinct rises can be distinguished. The first rise (L∗ to H-) begins at the low target of the pitch accent and subsequently flattens out. The second rise (H- to H%) begins at the onset [l] consonant of the final syllable. (Note how the elbow in the contour is approximately aligned with the location where the higher frequencies in the spectrogram are dampened from the [l].)

To summarize, the simple rise pattern in English involves a rise from the nuclear stressed syllable, hence the rise often spans multiple syllables. In contrast, a boundary rise in Japanese must be confined to the phrase-final syllable. In addition, both languages have double-rise patterns, where (for both languages) the second rise is confined to the final syllable. The distinguishing factor is the first rise. In Japanese, the first rise occurs over the first two moras of the phrase (thus marking its left boundary), whereas in English the first rise begins at the nuclear pitch accent. Schematically, these facts can be summarized as follows:

<table>
<thead>
<tr>
<th>Rise type</th>
<th>Language</th>
<th>Antepenultimate F0 target</th>
<th>Penultimate F0 target</th>
</tr>
</thead>
<tbody>
<tr>
<td>single</td>
<td>Japanese</td>
<td>accented syllable in final phrase</td>
<td>final syllable</td>
</tr>
<tr>
<td>single</td>
<td>English</td>
<td>at or before beginning of final phrase</td>
<td>nuclear pitch accent</td>
</tr>
<tr>
<td>double</td>
<td>Japanese</td>
<td>beginning of final phrase</td>
<td>final syllable</td>
</tr>
<tr>
<td>double</td>
<td>English</td>
<td>nuclear pitch accent</td>
<td>final syllable</td>
</tr>
</tbody>
</table>

*Table 4.1: Comparison of single and double boundary rises in English and Japanese *

In both cases, the phonological makeup of the boundary rise differs between languages. For example, in a Japanese simple rise contour, the beginning of the boundary rise would be a phrase-final L% boundary tone (which is often 'fused' with the low target at the end of a H*+L accent). In contrast, in an English simple rise contour, the beginning of the boundary rise may be
a L* pitch accent (which may be 'fused' with a L- phrase tone). Despite their different underlying phonological makeup, the two contours share a superficial similarity in that both are simplex boundary rises. As such, a Japanese EFL learner may treat the two as analogous (cf. the 'equivalence classification' of the Speech Learning Model). In such cases, from the learner's perspective, the relevant cross-linguistic difference would be one of [Realization, Alignment] in terms of the typology from Chapter 2.

As with the two other research questions discussed above, this could, in principle, lead to transfer in their L2 English production. The present dissertation focuses specifically on the case of single rises (i.e. a sequence of a fall plus a boundary rise). More specifically, the following research question is asked.

(8) **Research Question 3**: In their L2 English production of interrogative sentences ending in one or more unstressed syllables, how often do Japanese EFL learners produce a simple boundary rise from within the final syllable?

The facts sketched out above suggest that, in their L2 English production, Japanese EFL learners will frequently fall from the nuclear syllable and rise from the phrase-final syllable (i.e. the accented pattern in Figure 4.11). The rise is hypothesized to begin from the final syllable (e.g. the *lade* of *marmalade*), indicating the transfer of the rigid alignment of L% from the L1. In utterances with one or more post-nuclear syllables, this would be non-targetlike since, in pragmatically unmarked contexts, the boundary rise should begin from the nuclear syllable in native speakers' productions of the same sentences.

With the research questions thus established, the following chapter discusses the step-by-step procedure used to draw on corpus data in order to answer these questions.
Chapter 5: Procedure

This chapter details the procedure of selecting and processing the data for analysis (including the application of the method from Chapter 3). First, section 5.1 describes the selected L2 speech corpus. Section 5.2 then lists the sentences in this corpus that were chosen for examining each of the research questions. Next, section 5.3 discusses the rating data that is included with the corpus and how it can be used to gauge learners' proficiencies in L2 English. Finally, section 5.4 overviews the workflow by which the data was processed in the course of the analysis.

5.1 Description of the dataset

The discussion below first provides a broad description of the overall nature of the corpus (§5.1.1), followed by a more detailed overview of its structure and makeup (§5.1.2). Then, based on a discussion of the methods the corpus designers used to elicit utterances from the learners (§5.1.3), several criteria are spelled out for selecting the specific sub-corpus used for the present dissertation (§5.1.4).

5.1.1 Overview of corpus as a whole

The corpus used for the present dissertation is Nihonjin gakusei-ni yoru yomiage eigo onsei dētabēsu ("[E]nglish speech database [R]ead by [J]apanese students", or "ERJ"). More information about the corpus can be found at http://research.nii.ac.jp/src/en/UME-ERJ.html. As
indicated in the name of the corpus, all speech samples represent recordings of people reading English words and sentences aloud. More specifically, the corpus contains speech data for a total of 202 Japanese EFL learners (both undergraduate and graduate students), randomly sampled from one of twenty Japanese colleges and universities participating in the corpus construction process. The random sampling was intended to avoid sampling bias (e.g. recruiting only highly motivated advanced learners), thus the resulting corpus is claimed to be balanced in terms of proficiency (Minematsu et al. 2002a:3). "If only voluntary speakers are collected for the recording, the database shall contain only English speech samples of rather good speakers of English." By design, then, the L1 and L2 for all 202 learners is controlled for, thus creating a useful homogeneity within this portion of the corpus.

In addition, to serve as a baseline for interpreting the learners' performance, the corpus includes parallel data from 20 native speakers of American English living in Japan (8 of whom were English teachers). Thus, the ERJ is not only an L2 speech corpus but also has a small native-speaker subcomponent. The consent form for the native speakers collected various pieces of potentially sociolinguistically-relevant demographic information (gender, age, native language, early residential history, and parents' birthplaces), but none of this information is distributed with the corpus. They are described as all being native speakers of "General American" as defined by Giegerich (1992), from which it can only be concluded that they are neither from the "East" (Eastern New England and New York City) nor the "South" (Virginia, North Carolina, South Carolina, Georgia, Louisiana, and Texas). Given that there is some evidence for intonation-related differences among American English dialects (Arvaniti & Garding 2007, Clopper & Smiljanic 2011), it is unfortunate that specific details are unavailable
regarding the native speakers' linguistic backgrounds. However, intonational variation among these dialects is perhaps more subtle and less pervasive than variation in other dimensions such as vowel quality, hence the uncontrolled heterogeneity ('noise') this introduces into the native speaker data is not expected to be too very extensive.

Overall, the corpus contains 89,095 soundfiles (of which at least 500 are empty dummy files due to experimental error). The corpus was originally built to create a rich dataset that could be used to train the data-intensive statistical models underlying modern Computer Assisted Language Learning (CALL) applications. The corpus construction was funded through a 'Scientific Research on Priority Areas' grant-in-aid from the Japanese government's Ministry of Education, Culture, Sports, and Technology as part of the larger umbrella project "advanced [U]tilization of [M]ultimedia to promote higher [E]ducation reform", abbreviated UME (Nakagawa (2003)). The committee and working group consisted of seven researchers at five different Japanese universities, most of whom had backgrounds in information processing and electrical engineering. The construction of the corpus took place from 2000 to 2002 and was then released publicly in May 2007 through the Speech Resources Consortium at the National Institute of Informatics. For research purposes, a copy of the corpus can be obtained without fee through a simple application process.¹

¹ For further documentation about the corpus published in English, see Minematsu et al. (2002a). An earlier report on the project is Minematsu et al. (2001a). For documentation in Japanese, Minematsu et al. (2003b) is the final publication of the project and goes into the greatest detail. Earlier reports on the project are Minematsu et al. (2003a) and Minematsu et al. (2001b). Discussion of the rating component can be found in Minematsu et al. (2002b). Sample soundfiles and information on how to obtain it from the Speech Resources Consortium can be found at http://research.nii.ac.jp/src/en/UME-ERJ.html
The ERJ is a suitable choice of dataset for the present purposes for at least five reasons. (1) All utterances were produced in isolation, with no surrounding discourse context (e.g. a dialogue) and with no interlocutor, ensuring the data is free of several factors that can compromise F0 measurements, such as laughter or overlapping speech with an interlocutor. (2) The lexical content of every soundfile is pre-determined based on the corpus design, facilitating analysis. (3) Every learner in every soundfile is performing the same task (reading aloud), thus making the results for the three research questions more directly comparable. (4) With its immense size (89,000 files and over 200 learners), it is possible to explore a wide range of questions with a sufficiently large number of observations for statistical analysis. (5) The baseline data for native speakers can serve as a valuable counter-point to the analysis (e.g. ensuring the native speakers are not treated as if they were transferring L1 Japanese). Due to these numerous advantages, the ERJ was deemed to be a suitable choice of corpus for the present dissertation.

5.1.2 Corpus structure

The stimulus materials were designed to elicit a wide variety of structures - both segmental and suprasegmental - at both the word and the sentence level, thus forming a 2x2 of [Words/Sentences] x [Segmental/Suprasegmental]. The [Words, Segmental] component consists of 900 words that are either 'phonemically balanced' (i.e., each of the phonemes of English appears approximately the same number of times) or form minimal pairs. The [Words, Suprasegmental] component consists of 109 words with various stress patterns. The [Sentences, Segmental] component contains 592 sentences of three types: (A) phonemically balanced sentences adapted from the Texas Instruments - Massachusetts Institute of Technology (TIMIT)
Acoustic-Phonetic Speech Corpus (Garofolo et al. 1993), (B) sentences with phonotactically difficult sound sequences for Japanese EFL learners, and (C) sentences designed by actual English teachers in Japan for testing pronunciation. Finally, the [Sentences, Suprasegmental] component contains 215 sentences with various intonation and rhythm patterns. Across all four types, this totals to $900+109+592+215 = 1816$ different 'items' (i.e. words and sentences).

Further details about each of these sets of stimulus materials (including examples) are presented in the following table. (The final column is explained later below in §5.3.)

<table>
<thead>
<tr>
<th>Corpus component</th>
<th>Contents</th>
<th>Example(s)</th>
<th>Rated</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Words</strong></td>
<td>Segmental</td>
<td>300 phonemically-balanced words</td>
<td>rub, slip, smile, strife, such</td>
</tr>
<tr>
<td></td>
<td></td>
<td>600 words forming minimal pairs</td>
<td>luck/lack, rope/robe, sink/sing</td>
</tr>
<tr>
<td></td>
<td>Suprasegmental</td>
<td>109 words/phrases with various stress patterns</td>
<td>underestimate, broad-minded, life insurance policy</td>
</tr>
<tr>
<td><strong>Sentences</strong></td>
<td>Segmental</td>
<td>460 phonemically balanced sentences from TIMIT</td>
<td>Her classical repertoire gained critical acclaim.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>32 sentences with difficult phonotactic sequences</td>
<td>San Francisco is one-eighth as populous as New York.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100 sentences designed for testing pronunciation</td>
<td>This is thick and that is thin.</td>
</tr>
<tr>
<td></td>
<td>Suprasegmental</td>
<td>94 sentences with various intonation patterns</td>
<td>Cauliflower, broccoli, cabbage, sprouts, and onions.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>121 sentences with various rhythm patterns</td>
<td>Come to tea with John and Mary at ten.</td>
</tr>
</tbody>
</table>

*Table 5.1: Structure of ERJ corpus as a whole*

Given the large size of the set of materials (nearly two thousand words and sentences), the corpus designers deemed it impractical to have every speaker produce every item. Accordingly, the words were grouped into 5 lists and the sentences were grouped into 8 lists, and each speaker was randomly assigned one word list plus one sentence list. The native speakers read much larger subsets of the materials, perhaps partly due to the relative ease of reading many
sentences in one's L1 and partly due to the fact they were professional English teachers. More specifically, the materials were broken into two lists ('X' and 'Y'), with every component of the corpus (both words and sentences) split equally between the two lists. Nine native speakers read just list X, nine others read just list Y, and the final two read both lists (i.e. the entire set of materials).

The splitting of the materials into lists has the unfortunate consequence that the dataset is not fully crossed, i.e., it is not true that every speaker read every item, nor is it true that every item was read by every speaker. However, given the large number of speakers represented in the corpus, data on any given item is available from several dozen speakers (the exact number depending on various factors), both native and non-native. Likewise, given the large size of the materials, a substantial volume of speech data (both words and sentences) is available for any given speaker. As such, both within-subject and within-item analyses are both possible, partially mitigating the problem of lacking a fully-crossed design.

5.1.3 Elicitation methods and rationale

Recall from above that ultimate goal for the ERJ was to serve as a database for building better CALL programs. In order to be useful for this purpose, the learners' utterances must be of such a nature that they can be "treated adequately and correctly by the current speech technologies" (Minematsu et al. 2002a:2). Minematsu et al. argue that an optimal corpus for this purpose would be internally homogeneous, i.e. with minimal intra- and inter-learner variation. With such a corpus, the set of all learners as a whole would provide a clear and consistent data source for the speech processing algorithms as to the sort of interlanguage characteristics that are "found rather commonly and frequently in Japanese speaking of English" (Minematsu et al. 2002a:2). Minematsu et al. argue that an optimal corpus for this purpose would be internally homogeneous, i.e. with minimal intra- and inter-learner variation. With such a corpus, the set of all learners as a whole would provide a clear and consistent data source for the speech processing algorithms as to the sort of interlanguage characteristics that are "found rather commonly and frequently in Japanese speaking of English" (Minematsu et al. 2002a:2).
Thus, the corpus was not designed to investigate individual differences but rather level population-level trends.

Toward this end, the corpus designers sought to avoid recording speech patterns that are uncharacteristic for a particular learner (e.g. a disfluency due to anxiety from being recorded) and, instead, wanted the obtained data to be an accurate reflection of what each learner truly believed to be the correct pronunciation (Minematsu et al. 2002a:3-4). As such, the learners were given their lists of stimuli in the days prior to the recording session so they could practice and rehearse as much as they liked. Moreover, during the recording session itself, they could produce a given item multiple times until they felt confident and comfortable with it. Both of these methodological decisions were intended to reduce within-learner variation by making each learner more internally consistent.

In addition, the corpus designers sought to ensure all learners were aware of what the intended target (i.e. native English) norm was for every item. The intention was, for example, to prevent learners from merely guessing at how to produce a word they had never encountered before based on its spelling. Also, since a given sentence can have various prosodic realizations depending on the desired interpretation, clarifying the target norm in this way would prevent different learners from reading the same sentence multiple ways merely because they were unsure of which 'reading' the researchers had in mind. By excluding cases like these, a learner's divergences from target norms could be more directly attributed to their phonological knowledge rather than other external factors.

The specific way the corpus designers make clear the intended target for each word and sentence was through the use of metalinguistic annotations written on the page - always on the
sheet used for practicing and, for certain components of the overall corpus, also on the sheet used during the actual recording. The annotations took the form of ASCII phoneme codes modified from TIMIT (Garofolo et al. 1993), word segmentation (in the form of square brackets), stress marks, and/or arrows for intonation. Here again, this depended on the component of the corpus being tested, with the annotations emphasizing what was most relevant for each specific elicitation task. An example is given below (one of the 121 sentences with various rhythm patterns):

(1) Come to tea with John and Mary.
    / +    - @ / -    +    - @ - /
    [K AH1 M] [T UW1] [T IY1] [W IH1 DH] [JH AA1 N] [AE1 N D] [M EH1 R IY0]

The second line in this example indicates the prosodic structure of the utterance. Slashes ("/") indicate prosodic boundaries, at marks ("@") indicate primary stresses, plus signs ("+") indicate secondary stresses, and minus signs ("-")) indicate unstressed syllables. The third line in this example contains the phoneme codes for the utterance, consisting of 1-2 letter symbols for each segment (e.g. "JH" for [ʤ] or "IY" for [i]). In addition, vowels have a number suffixed at the end to indicate stress level (primary=1, secondary=2, unstressed=0). For example, in the final word "Mary", the first vowel ("EH") is suffixed with "1" whereas the second vowel ("IY") is suffixed with "0", indicating the word's trochaic (strong-weak) status.

Learners were explicitly taught the meanings of these codes. Not only was an equivalency table (with entries like "B for bay (B EY1)") provided on the information sheet distributed to subjects, but a website was created with downloadable files so they could listen to a native speaker pronounce each word in the table (Minematsu et al. 2003b:262). Thus, through
these two methods for studying at home (presumably in the 24 hours preceding the recording session), learners could be expected to be sufficiently familiar with the codes.

5.1.4 Selected sub-corpus

For a study on L2 intonation, the "94 sentences with various intonation patterns" component of the corpus would seem a natural choice. This component consisted of the sentence sets in the following table, labeled (under "Type") with direct translations from the corpus's documentation. (The parenthesized number in the last column indicates contextualizing sentences, e.g. where learners were recorded reading *Is it John who writes poetry?* before the target sentence *No, it is Bill who writes poetry.*

<table>
<thead>
<tr>
<th>Type</th>
<th>Example</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phrase-level stress</td>
<td><em>take care of themselves</em></td>
<td>20</td>
</tr>
<tr>
<td>Chunking-related intonation differences</td>
<td><em>I don't know(.) Miss Brown.</em></td>
<td>16</td>
</tr>
<tr>
<td>Intonation when enumerating (rising or flat)</td>
<td><em>Cauliflower, broccoli, cabbage, sprouts, and onions.</em></td>
<td>4</td>
</tr>
<tr>
<td>Classification by tone</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rising tone</td>
<td><em>When I came, he greeted me warmly.</em></td>
<td>2</td>
</tr>
<tr>
<td>Falling-rising tone</td>
<td><em>I should go. [But I don't think I will.]</em></td>
<td>2</td>
</tr>
<tr>
<td>Rising-falling tone</td>
<td><em>Who knows? [= Nobody knows.]</em></td>
<td>6</td>
</tr>
<tr>
<td>Flat tone</td>
<td><em>He drank, he stole, he was soon despised.</em></td>
<td>1</td>
</tr>
<tr>
<td>Classification by sentence pattern</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Declarative</td>
<td><em>Legumes are a good source of vitamins.</em></td>
<td>(2+)3</td>
</tr>
<tr>
<td>Yes/no interrogative</td>
<td><em>Are legumes a good source of vitamins?</em></td>
<td>(1+)5</td>
</tr>
<tr>
<td>Alternative interrogative</td>
<td><em>Is this elevator going up or down?</em></td>
<td>4</td>
</tr>
<tr>
<td>Wh interrogative</td>
<td><em>How long have you been waiting?</em></td>
<td>(3+)6</td>
</tr>
<tr>
<td>Tag question</td>
<td><em>I haven't seen you before, have I?</em></td>
<td>3</td>
</tr>
<tr>
<td>Exclamative</td>
<td><em>Isn't it wonderful weather!</em></td>
<td>2</td>
</tr>
<tr>
<td>Differences based on old vs. new information</td>
<td><em>No, it is Bill who writes poetry.</em></td>
<td>(7+)7</td>
</tr>
</tbody>
</table>

| Total:                                                 | (13+)81                                                                |        |

Table 5.2: *Structure of materials in the ERJ's 'Sentence, Intonation' component*

While such materials are attractive for tapping learners' knowledge of intonation, the metalinguistic annotations that accompanied these sentences are problematic for the present
study's purposes. In particular, these sentences were accompanied by arrows indicating where one's pitch should rise or fall, which were shown during practice as well as the actual recording. While, as described above, such annotations were well-suited to the corpus's original purposes, they are inappropriate for a study on naturalistic interlanguage development. The core problem is that, with such annotations, it is impossible to know exactly how much of a given utterance produced by a learner reflect their implicit phonological knowledge and how much was explicit knowledge 'learned' on the spot from rehearsing with the metalinguistic annotations on the page (and is therefore merely an artifactual byproduct of the elicitation process itself). Since the goal of the present study is to understand each learner's phonological knowledge, it is of greatest interest what an L2 learner would do in normal circumstances, i.e. without any such metalinguistic annotations.

Due to these considerations, another part of the overall corpus needed to be selected. Since similar remarks apply to the "121 sentences with various rhythm patterns" (namely the use of rhythm marks), and since rhythm is intertwined with intonation in various ways, the entire [Sentences, Suprasegmental] component of the corpus was thus rendered unusable. Moreover, the [Words, Segmental] and [Words, Suprasegmental] components to the corpus are a poor fit because the focus largely on individual words whereas the present study is concerned with utterance-level intonation. These criteria left only the [Sentences, Segmental] component of the corpus, which was ultimately adopted for as the dataset for the present study.

Recall from above that this subset consists of 592 sentences with three sub-components: 460 phonemically balanced sentences adapted from the TIMIT corpus, 32 sentences with phonotactic sequences that are difficult for Japanese EFL learners, and 100 sentences designed
by actual English teachers in Japan for testing pronunciation. Since the present study is concerned with intonation, not segmentals, the three-way contrast between these different subtypes of sentence is not relevant for the present study. As such, throughout the rest of the present dissertation, the distinction between these three subgroups is collapsed (such that all three are treated together as a single 592-sentence whole).

Since this component of the corpus was focused on segmentals, during the practice phase, in addition to the normal English orthography for each sentence, learners saw the ASCII phoneme codes modified from TIMIT, segmented into words by square brackets. (An example of this sort of representation can be seen in last line of (1) above.) However, Minematsu et al. (2002a:3) reasoned that "reading sentences with referring to phonemic symbols is expected to induce unnatural pronunciation"; more specifically, "[w]ith the phonemic symbols for each word, some speakers may not read a sentence but a sequence of isolated words." As such, no phonetic symbols were shown during the actual recording session, i.e., every sentence was elicited with only its normal representation in English orthography.

This particular method of elicitation is not particularly problematic for the purposes of the present study. Given the markedly foreign appearance of the TIMIT transcriptions, many learners may not have even looked closely at the phonetic symbols when practicing the stimuli, and even the learners who did may not have gained much from it (especially since the transcription was unavailable during the actual recording itself). Even if a learner was influenced by the transcriptions, since they contain segmental information (plus stress levels), this should

---

2 Two of the 100 "sentences" designed by English teachers for testing pronunciation are technically sequences of two sentences (produced by the same learner in sequence). These are the following: (1) "Where do you live? I live in the woods." and (2) "I counted the dogs. There were ten of them."
mostly influence their production of segmentals. In fact, since the learner's explicit attention was
drawn to something other than prosody during the recording of these sentences, learners'
production of these sentences can be thought of as tapping each learner's 'unguarded' implicit
knowledge of prosody (i.e. relatively uninfluenced by conscious metalinguistic awareness). In
this sense, these stimuli can actually be thought of as ideal for a study of intonation.

What is crucial for the purposes of the present study is that, for these sentences, learners
were never given any metalinguistic annotations regarding intonation (e.g. curved arrows
pointing up or down). Of course, the learners saw such annotations when producing their piece
of the "94 sentences with various intonation patterns" component of the corpus (both during the
practice as well as the actual recording). However, it is unlikely that this kind of brief exposure
to arrows printed on the page would have a sufficiently deep impact on the learners as to
generalize to the other sentences elsewhere in a given student's list, especially with their attention
drawn to segmentals. Thus, the method of elicitation for the selected set of 592 sentences is at
least compatible with - if not ideal for - the purposes of the present study.

Recall from above that the words were grouped into 5 lists and the sentences were
grouped into 8 lists, and each learner was randomly assigned one word list plus one sentence list.
Each of the eight sentence lists contained different (non-overlapping) subsets of the selected 592
sentences, ranging in size from 57 to 78 sentences each. The number of learners reading each of
the eight lists ranged from 22 to 28 learners. (No individual learner read more than one list.) This
means there are 22-28 data points from Japanese EFL learners for any given sentence. The full
details about the number of sentences and the number of learners for each list are provided
below.
### Table 5.3: Number of sentences and learners for each of the eight lists

<table>
<thead>
<tr>
<th></th>
<th>Number of sentences</th>
<th>Number of learners</th>
</tr>
</thead>
<tbody>
<tr>
<td>List 1</td>
<td>76</td>
<td>26</td>
</tr>
<tr>
<td>List 2</td>
<td>76</td>
<td>25</td>
</tr>
<tr>
<td>List 3</td>
<td>76</td>
<td>22</td>
</tr>
<tr>
<td>List 4</td>
<td>76</td>
<td>23</td>
</tr>
<tr>
<td>List 5</td>
<td>78</td>
<td>26</td>
</tr>
<tr>
<td>List 6</td>
<td>78</td>
<td>27</td>
</tr>
<tr>
<td>List 7</td>
<td>78</td>
<td>25</td>
</tr>
<tr>
<td>List 8</td>
<td>57</td>
<td>28</td>
</tr>
<tr>
<td><strong>Total:</strong></td>
<td><strong>592</strong></td>
<td><strong>202</strong></td>
</tr>
</tbody>
</table>

The structure for the native speakers was much simpler. Since the entire set of materials were split into two equal-sized halves (list 'X' and list 'Y') for these speakers, each of the two lists had half (296) of the 592 sentences. The number of speakers also matches that of the overall corpus: 9 native speakers reading list X, 9 reading list Y, and 2 reading both. This means that, for any given sentence (from either list X or list Y), there are always exactly 11 native-speaker productions to serve as the control.

### 5.2 Selection of sentences

Ideally, all productions of all 592 sentences by all native speakers and all learners would be analyzed. However, this is impractical, as it totals to 12,898 tokens for learners and 6,512 tokens for native speakers. Consequently, an alternate approach was deemed necessary. The avenue pursued here is to select, for each of the three research questions, a separate set of sentences where that research question can be most fruitfully examined. Each research question can be thought of as picking out a certain kind of text that produces data of the right type in order to answer that particular research question. Thus, if the relevant characteristic for a given research question is absent from a particular sentence (e.g. a boundary rise for Research Question...
3), then that sentence is excluded. Since the inferences about each of these subsets will be used to make inferences that generalize to all sentences as a whole, each sentence set should be a thorough and accurate representation of the phenomenon in question.

The sentence sets are chosen based on two additional criteria. First, the three sets of sentences must not overlap. This ensures that the answers to the different research questions are justified based on fully independent grounds. It also means that a larger portion of the overall corpus will be annotated (making generalization safer than if the entire analysis was based on a small set of the same sentences). Second, the four sets of sentences must be roughly equal-sized. This is necessary because a convincing answer to the higher-level question about the hierarchy of frequency is only possible with balanced cells, creating an 'equal chance' for the three different types transfer to appear.

In order to ensure the learners were familiar with the stimulus words making up the sentences in these sets, it would have been possible to constrain the words to be of high frequency. Indeed, as will be seen in the sections below, several words (including those in target positions) are somewhat advanced vocabulary, e.g. *ambidextrous* or *pizzerias* for Research Question 1. However, there are at least four problems with such an approach: (1) Doing so would require selecting some cutoff point between "low frequency" and "high frequency" words, which is arbitrary. (2) Frequency statistics depend to some extent on the corpus selected. (3) Frequency itself only indirectly indexes what is actually relevant - the familiarity of the words to these specific learners. (4) This would have made it impossible to collect balanced sets of 14 sentences for all three research questions. Most importantly, though, such a move may not be necessary in the first place. As discussed above in §5.1.3, learners were given the materials well in advance
and study it as much as they liked before the recording session. As such, at least some of the
learners presumably used a dictionary to look up the low-frequency words they did not know (if
anything, to spare them the embarrassment of mispronouncing a word in front of the
experimenter). While there were undoubtedly several students who came into the recording
session lacking familiarity with some of the low-frequency words therein, many (if not most)
learners presumably comprehended the meaning of the sentences they were reading. As such,
construct-irrelevant noise stemming from a lack of familiarity with the target words should be of
relatively small magnitude. Thus, most of the observed non-targetlikeness should be from the
phenomenon of interest - crosslinguistic transfer.

Based on the above considerations, three non-overlapping sets of 14 sentences were
chosen to explore the three research questions (since 14 is the largest number for which this is
possible). The exact sentences for Research Questions 1 through 3 are laid out in the first three
sections below (§5.2.1 through §5.2.3).

5.2.1 Sentences selected for Research Question 1

Research Question 1 seeks evidence for an L1-transferred phrasal H- in Japanese EFL
learners' production. The ideal sentence for investigating this research question should meet three
criteria. First, recall from §4.1.1 that Japanese phrasal H- occurs on the first or second mora
within a given phrase. Thus, the relevant position for this research question is at the very
beginning of a prosodic phrase. Given how the same sentence can be read in various ways (as
afforded by the flexibility of the English intonation system), and given the wide range of
variability expected of L2 learners (in not only their proficiency in intonation but also in
syntactic parsing), it is hard to predict exactly where prosodic phrases will begin in the middle of
an utterance. In contrast, utterance-initial position is effectively guaranteed to be the beginning
of a prosodic phrase (being inherently post-pausal). Since this is precisely the environment for
phrasal H-, the first criterion for Research Question 1 is that the environment for the phrasal H-
must be in utterance-initial position.

Second, the string of syllables in this position should be all one word and not, for
example, a sequence of a function word (such as the) followed by a content word. The latter
introduces the possibility that the learner may introduce a prosodic break between the two words
in question (e.g. due to a disfluency or processing constraints), which would complicate the
environment for the appearance of the phrasal H-. Consistently constraining all sentences to
begin with a single target word with the right kind of phonological environment circumvents this
potential problem. Thus, the second criterion is that the environment for the phrasal H- should be
confined to a single word.

Third, recall from §4.1.1 that the phrasal H- occurs before the first pitch accent.
Generally speaking, evidence for a given tone is clearest when it is far removed from
neighboring tones, i.e. if multiple intervening syllables prevent 'tonal crowding' from obscuring
the patterns in the F0 track. Thus, the ideal context would be one where the potential phrasal H-
position is well removed from the first syllable in the phrase bearing a pitch accent (i.e., in the
case of English, a word's primary stress). Thus, the third criterion is that the primary stress for
the target word should be at least two syllables away from the beginning of the utterance.

An example of a sentence meeting all three of these criteria is Curiosity and mediocrity
seldom coexist. Regardless of how the middle portion of this sentence is phrased, curiosity will
always be a single word in utterance-initial position, thus satisfying the first two criteria.
Moreover, the primary stress is on the third syllable into the word (cù.ri.ó.si.ty), thus satisfying the third criterion. Since the initial syllable in this word (cù-) is light, the expected location for the phrasal H- is the second mora (on -rí-). Thus, if high target is visible there in a learner’s utterance, it would be possible to attribute it to an L1-transferred phrasal H-.

Using these two criteria, the following 14 sentences were selected for the purposes of Research Question 1:

<table>
<thead>
<tr>
<th>#σ</th>
<th>First word</th>
<th>Rest of sentence</th>
<th>List ID</th>
<th>Stimulus ID</th>
<th>Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Em.plòy.ée</td>
<td>layoffs coincided with the company's reorganization.</td>
<td>S3_049</td>
<td>S_PH_B_1_169</td>
<td>22+11</td>
</tr>
<tr>
<td></td>
<td>Màs.quiz.áde</td>
<td>parties tax one's imagination.</td>
<td>S2_036</td>
<td>S_PH_B_1_096</td>
<td>25+11</td>
</tr>
<tr>
<td></td>
<td>Cà.ta.stró.phic</td>
<td>economic cutbacks neglect the poor.</td>
<td>S1_050</td>
<td>S_PH_B_1_050</td>
<td>26+11</td>
</tr>
<tr>
<td></td>
<td>Am.bi.áex.trous</td>
<td>pickpockets accomplish more.</td>
<td>S1_051</td>
<td>S_PH_B_1_051</td>
<td>26+11</td>
</tr>
<tr>
<td></td>
<td>À.ca.dé.mic</td>
<td>aptitude guarantees your diploma.</td>
<td>S1_056</td>
<td>S_PH_B_1_056</td>
<td>26+11</td>
</tr>
<tr>
<td></td>
<td>Còn.ti.nén.tal</td>
<td>drift is a geological theory.</td>
<td>S2_003</td>
<td>S_PH_B_1_063</td>
<td>25+11</td>
</tr>
<tr>
<td></td>
<td>Àr.ti.íf.ial</td>
<td>intelligence is for real.</td>
<td>S3_008</td>
<td>S_PH_B_1_128</td>
<td>21+11</td>
</tr>
<tr>
<td></td>
<td>Pí.zze.rí.as</td>
<td>are convenient for a quick lunch.</td>
<td>S3_015</td>
<td>S_PH_B_1_135</td>
<td>22+11</td>
</tr>
<tr>
<td></td>
<td>Scí.en.ti.fic</td>
<td>progress comes from the development of new techniques.</td>
<td>S3_018</td>
<td>S_PH_B_1_138</td>
<td>21+11</td>
</tr>
<tr>
<td>3</td>
<td>Àg.ri.cúl.tu.ral</td>
<td>products are unevenly distributed.</td>
<td>S3_022</td>
<td>S_PH_B_1_142</td>
<td>22+11</td>
</tr>
<tr>
<td></td>
<td>Cù.ri.ó.si.ty</td>
<td>and mediocrity seldom coexist.</td>
<td>S3_046</td>
<td>S_PH_B_1_166</td>
<td>22+11</td>
</tr>
<tr>
<td></td>
<td>Vi.et.na.mése</td>
<td>cuisine is exquisite.</td>
<td>S3_032</td>
<td>S_PH_B_1_152</td>
<td>22+11</td>
</tr>
<tr>
<td></td>
<td>Co.ò.per.á.ión</td>
<td>along with understanding alleviate dispute.</td>
<td>S2_024</td>
<td>S_PH_B_1_084</td>
<td>25+11</td>
</tr>
<tr>
<td></td>
<td>En.çy.cló.pé.di.as</td>
<td>seldom present anecdotal evidence.</td>
<td>S3_004</td>
<td>S_PH_B_1_124</td>
<td>22+11</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>327+154</td>
</tr>
</tbody>
</table>

Table 5.4: 14 sentences for Research Question 1, grouped by number of syllables before primary stress

The second column of the table ("First word") contains the first word of each sentence (the target of the analyses for Research Question 1). The remaining portion of each sentence is presented in the third column ("Rest of sentence"). The first word is presented in normal English orthography but is segmented into syllables using periods as a delimiter. Moreover, stressed
syllables are indicated both with an acute/grave accent over the relevant letter as well as formatting (underline=secondary stress, bold=primary stress). The first column ("#σ") shows the number of syllables in the word before the primary stress, demonstrating that indeed it occurs in all words at least two syllables from the beginning of the word, as required by the second criterion discussed above. Moreover, many of these sentences begin with compound nouns or adjective + noun sequences, where the first pitch accent may possibly occur much later (e.g. *catastrophic economic cutbacks*). In such cases, the phrasal H- would be even further removed from the pitch accent, making its appearance all the more clearly distinguishable.

The "List ID" and "Stimulus ID" columns contain the code for each sentence used within the ERJ, provided here for reference purpose only. (The list number (1 through 8) is the second digit inside the List ID.) The final column "Tokens" indicates how many tokens are available for each sentence (as determined by the list it falls in). The first number indicates the number for the learners (somewhere between 21 and 26) and the second number (always 11) indicates the number for the native speakers. ³ In sum, there are 327 learner tokens and 154 native speaker tokens, hence 481 total soundfiles (with an aggregate length of 34.95 minutes).

### 5.2.2 Sentences selected for Research Question 2

Research Question 2 seeks evidence for an L1-transferred phrasal L% in Japanese EFL learners' production, such that learners were expected to use this tone to break larger utterances into more smaller prosodic phrases than English native speakers would (using L-). Two criteria were used to select sentences. First, a sentence could not have any 'inherent' prosodic breaks (e.g. *Artificial...* and *Scientific...*) because two of the relevant recordings (JE/TUT/F04/S3_008.wav and JE/TEI/F02/S3_018.wav) are empty dummy files.

³ There are only 21 learner tokens for two of the sentences ("Artificial..." and "Scientific...") because two of the relevant recordings (JE/TUT/F04/S3_008.wav and JE/TEI/F02/S3_018.wav) are empty dummy files.
with a comma plus 'but' or 'and'). Such things would effectively guarantee breaks to occur, whereas the research question focuses on 'normal' breaks introduced by the speaker utterance-medially. Second, the ideal sentence for investigating this research question should be many words long. In theory, each pair of adjacent words is a potential boundary location (with, of course, some more probable than others), hence longer sentences increase the chances of a speaker inserting a prosodic boundary. Sentences meeting both of these criteria generally contained long stretches of material that native speakers would group together into one phrase but a learner might parse into several smaller phrases. For example, in "I caught a strange insect in the inmost part of the forest", a native speaker might parse this into two phrases, as in (2a), whereas a learner might parse it into four (or more) phrases, as in (2b).

(2) a. {I caught a strange insect} {in the inmost part of the forest}
   b. {I cáught} {a strange insect} {in the inmost párt} {of the fórest}

All of the following 14 sentences were selected for the purposes of Research Question 2 since they lack 'inherent' breaks, are 10+ words long, and can be expected to exhibit a kind of learner-vs.-native discrepancy like the one in (2). If length is counted in syllables rather than words, the sentences range from 11 syllables (He picked up nine pairs of socks for each brother.) to 23 syllables (Al received a joint appointment in the biology and the engineering departments.) with a median of 15.4

---

4 Due to experimental error, the indefinite article 'a' was omitted from the sentence "I caught a strange insect in the inmost part of the forest" in an early version of the experimental materials presented to a small minority of the learners and the natives.
<table>
<thead>
<tr>
<th>#w</th>
<th>Text</th>
<th>List ID</th>
<th>Stimulus ID</th>
<th>Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>He picked up nine pairs of socks for each brother.</td>
<td>S7_037</td>
<td>S_PH_B_1_397</td>
<td>25+11</td>
</tr>
<tr>
<td></td>
<td>Ralph prepared red snapper with fresh lemon sauce for dinner.</td>
<td>S8_036</td>
<td>S_PH_B_1_456</td>
<td>28+11</td>
</tr>
<tr>
<td>11</td>
<td>Right now may not be the best time for business mergers.</td>
<td>S7_016</td>
<td>S_PH_B_1_376</td>
<td>25+11</td>
</tr>
<tr>
<td></td>
<td>The rich should invest in black zircons instead of stylish shoes.</td>
<td>S8_022</td>
<td>S_PH_B_1_442</td>
<td>28+11</td>
</tr>
<tr>
<td></td>
<td>Shell shock caused by shrapnel is sometimes cured through group therapy.</td>
<td>S8_025</td>
<td>S_PH_B_1_445</td>
<td>28+11</td>
</tr>
<tr>
<td></td>
<td>I caught a strange insect in the inmost part of the forest.</td>
<td>S3_075</td>
<td>S_PH_E_1_035</td>
<td>22+11</td>
</tr>
<tr>
<td></td>
<td>His failure to open the store by eight cost him his job.</td>
<td>S4_056</td>
<td>S_PH_B_1_236</td>
<td>23+11</td>
</tr>
<tr>
<td></td>
<td>Those answers will be straightforward if you think them through carefully first.</td>
<td>S6_017</td>
<td>S_PH_B_1_317</td>
<td>27+11</td>
</tr>
<tr>
<td></td>
<td>Al received a joint appointment in the biology and the engineering departments.</td>
<td>S6_032</td>
<td>S_PH_B_1_332</td>
<td>27+11</td>
</tr>
<tr>
<td></td>
<td>We revised the original plan because of the lack of the budget.</td>
<td>S7_009</td>
<td>S_PH_E_1_084</td>
<td>25+11</td>
</tr>
<tr>
<td></td>
<td>The football team coach has a watch as thin as a dime.</td>
<td>S8_006</td>
<td>S_PH_B_1_426</td>
<td>28+11</td>
</tr>
<tr>
<td></td>
<td>Gus saw pine trees and redwoods on his walk through Sequoia National Forest.</td>
<td>S6_040</td>
<td>S_PH_B_1_340</td>
<td>27+11</td>
</tr>
<tr>
<td></td>
<td>Each untimely income loss coincided with the breakdown of a heating system part.</td>
<td>S7_009</td>
<td>S_PH_B_1_369</td>
<td>24+11</td>
</tr>
<tr>
<td>15</td>
<td>These commonwealths will not long bear a state of subjection to the republic of Paris.</td>
<td>S5_062</td>
<td>S_PH_D_1_018</td>
<td>26+11</td>
</tr>
</tbody>
</table>

**Total:** 363+154

**Table 5.5: 14 sentences for Research Question 2, grouped by number of words in sentence**

The number of words in each sentence is displayed in the first column ("#w"). Since the number of words per sentences ranges from 10 to 15 ('N'), this means there are 9 to 14 ('N-1') potential prosodic phrase boundary sites per sentence. The second column "Text" contains the full text for each sentence. The remaining columns are the same as those of Table 5.4 above (for Research Question 1). "List ID" and "Stimulus ID" contain the ERJ's codes for each sentence, and "Tokens" contains the number of soundfiles available for each sentence. The first number indicates the number for the learners (somewhere between 22 and 28) and the second number
(always 11) indicates the number for the native speakers. In sum, there are 363 learner tokens and 154 native speaker tokens, hence 517 total soundfiles (with an aggregate length of 47.68 minutes).

5.2.3 Sentences selected for Research Question 3

Research Question 3 began with the observation that, in a boundary rise, the Japanese boundary L% exhibits a rigid alignment, obligatorily aligned to the phrase-final syllable. This stands in contrast to English, where many boundary rises can begin earlier in the utterance (e.g. from a L* pitch accent). This cross-linguistic difference led to the hypothesis that the low target in Japanese EFL learners' boundary rises should frequently be aligned to the phrase-final syllable. The following research question was formalized to test this prediction:

(3) **Research Question 3**: When Japanese EFL learners produce boundary rises at the end of English utterances ending in one or more unstressed syllables, what is the shape of the contour over the post-nuclear material? In particular, where does the boundary rise begin?

A given sentence must meet two criteria in order to be usable to explore this research question. First, in the speech of a native speaker, each sentence must obligatorily be produced with a boundary rise in final position. Of course, a boundary rise must be present in the first

---

5 Three of the tokens in the selected set require special mention. First, as with Research Question 1, one token (JE/TUT/M07/S7_009.wav) was excluded for being an empty dummy file. Thus, only 24 tokens are available for the relevant sentence ("Each untimely...") even though the sentences in list 7 usually have 25 sentences. Second, in file JE/TOH/M06/S6_032.wav, the learner is heard saying the number thirty-two in Japanese before beginning the sentence ("Al received..."). It appears learners read aloud a running index number before saying each sentence, and while those were normally cut out from the final version of the audio files, this was accidentally retained in this particular recording. For the purposes of analysis, this extra word was ignored and treated as if it were silence. Third and finally, file JE/RT/HF6/S8_025.wav is over 52 seconds long, containing a raw uncut recording of the learner starting and re-starting the target sentence ("Shell shock...") several times due to disfluencies as well as some renditions of the following sentence in the elicitation scheme (S8_026.wav). From this larger recording, the disfluency-free third production of the target sentence was extracted out and saved as a separate file for analysis.

6 In principle, boundary rises can occur both utterance-medially and utterance-finally. However, as noted above in §5.2.1, it is difficult to identify *a priori* where utterance-medial boundary rises will occur (especially given that they
place in order to analyze how its constituent targets are aligned. Yes-no (i.e. 'polar') questions are a suitable way to meet this criterion since a boundary rise is more or less required for this kind of utterance in both languages in question (more specifically, a fall-rise for Japanese and a simple rise for American English). (While productions of yes-no questions without boundary rises are technically possible in both languages, these are pragmatically marked and should not occur frequently in read-aloud speech.) Second, as stated in the research question itself, each sentence must end in one or more unstressed syllables. This simply translates to the restriction that the nuclear pitch accent must not fall on the utterance-final syllable - the case in which the English-like and Japanese-like patterns would overlap.

Since each of the following fourteen sentences from the ERJ stimulus materials meet both of these criteria, they were chosen for the purpose of exploring Research Question 3. Taking the last row of the table as an example, a native speaker might be expected to produce a (simplex) boundary rise from the syllable pho- in photograph, whereas a Japanese EFL learner may rise instead from -graph.

---

are often used to keep the conversational floor, a function that is inert for the present corpus). Thus, only utterance-final boundary rises are considered for this research question.
<table>
<thead>
<tr>
<th>#σ</th>
<th>Text (with final word syllabified)</th>
<th>List ID</th>
<th>Stimulus ID</th>
<th>Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Will Robin wear a yellow [li-ly]?</td>
<td>S1_012</td>
<td>S_P_B_1_012</td>
<td>26+11</td>
</tr>
<tr>
<td></td>
<td>Did dad do academic [bid-ling]?</td>
<td>S1_027</td>
<td>S_P_B_1_027</td>
<td>26+11</td>
</tr>
<tr>
<td></td>
<td>Does creole cooking use [cur-ry]?</td>
<td>S3_060</td>
<td>S_P_B_1_180</td>
<td>21+11</td>
</tr>
<tr>
<td></td>
<td>Do you hear the sleigh bells [ring-ing]?</td>
<td>S4_031</td>
<td>S_P_B_1_211</td>
<td>23+11</td>
</tr>
<tr>
<td></td>
<td>Are you looking for [em-ploy-ment]?</td>
<td>S4_053</td>
<td>S_P_B_1_233</td>
<td>23+11</td>
</tr>
<tr>
<td></td>
<td>May I order a parfait after I eat [din-ner]</td>
<td>S5_007</td>
<td>S_P_B_1_247</td>
<td>26+11</td>
</tr>
<tr>
<td></td>
<td>Do you have the yellow ointment [rea-dy]?</td>
<td>S5_026</td>
<td>S_P_B_1_266</td>
<td>26+11</td>
</tr>
<tr>
<td></td>
<td>Can the agency overthrow alien [for-ces]?</td>
<td>S5_027</td>
<td>S_P_B_1_267</td>
<td>26+11</td>
</tr>
<tr>
<td></td>
<td>Is he a [mail-man]?</td>
<td>S5_072</td>
<td>S_P_E_1_056</td>
<td>26+11</td>
</tr>
<tr>
<td></td>
<td>Will you please confirm government policy regarding waste [re-mo-val]?</td>
<td>S7_011</td>
<td>S_P_B_1_371</td>
<td>25+11</td>
</tr>
<tr>
<td>2</td>
<td>Did you eat lunch [yes-ter-day]?</td>
<td>S2_010</td>
<td>S_P_B_1_070</td>
<td>25+11</td>
</tr>
<tr>
<td></td>
<td>Did you buy any corduroy [o-ver-alls]?</td>
<td>S4_009</td>
<td>S_P_B_1_189</td>
<td>23+11</td>
</tr>
<tr>
<td></td>
<td>Would you allow acts of [vi-o-lence]?</td>
<td>S4_019</td>
<td>S_P_B_1_199</td>
<td>23+11</td>
</tr>
<tr>
<td></td>
<td>Do you know this man in this [pho-to-graph]?</td>
<td>S8_053</td>
<td>S_P_E_1_096</td>
<td>28+11</td>
</tr>
</tbody>
</table>

Total: 347+154

Table 5.6: 14 sentences for Research Question 3, grouped by number of post-nuclear syllables

In this table, the second column contains the text of each sentence. For the final word in each sentence (indicated with brackets), syllable boundaries are indicated with dashes and the stressed syllable therein is indicated with underlining and bolding. The first column indicates the number ("#") of unstressed syllables ("σ") after this stressed syllable. As indicated in the table, all sentences have either 1 or 2 such syllables, ensuring the availability of post-nuclear material. The third column in the table contains the 'List ID', i.e. the code for each sentence within the list structure of the stimulus materials. (The list number (1 through 8) is the second digit inside the List ID.) The fourth column is the code for each sentence within the design structure of the materials. The final column indicates how many tokens are available for each sentence (as determined by the list it falls in). The first number indicates the number for the learners (somewhere between 21 and 28) and the second number (always 11) indicates the number for the
native speakers.\textsuperscript{7} In sum, there are 347 learner tokens and 154 native speaker tokens, hence 501 total soundfiles (with an aggregate length of 26.75 minutes).

Two points are worth noting about the pitch accent location in these sentences. For many of the 14 sentences, placing the nuclear pitch accent on the sentence-final word would be pragmatically marked. For example, \textit{Do you hear the sleigh bells ringing?} would normally be produced with the nuclear stress on \textit{sleigh}, not \textit{ringing} (where it would change the sentence's meaning to one of contrastive focus). Given that the Japanese EFL learners in the present study are at various stages of interlanguage development, however, it is conceivable that many learners might place the nuclear pitch accent on the utterance-final word nonetheless. Thus, the selected sentences effectively ensure the availability of post-nuclear material even in the 'worst case scenario' that a learner puts the nuclear pitch accent on the final word (e.g. \textit{ringing} in the above example). If the nuclear pitch accent comes earlier, this only helps the situation by creating more post-nuclear material for the English-like and Japanese-like patterns to be more clearly distinguished.

Secondly, the final word in many of the sentences in the table have been borrowed into Japanese as loanwords (or could be adapted equivalently). Accordingly, it is possible that a Japanese EFL learner would produce these words with a stress location that matches the loanword's accent location rather than the actual English stress location. This creates an interpretative problem in two cases:

\textsuperscript{7} Despite being in List 3, which normally has 22 learners per sentence, the sentence "Does creole cooking use curry?" has only 21 learner tokens because one file (JE/TUT/M03/S3_060) is a dummy empty file.
(4) a. Japanese おばおる (zu) → English *'overall(s)'
   b. Japanese ふぉぐらふ → English *'photograph'

The problem is that, for example, if a learner produces a boundary rise from '-graph' in *photograph*, it could represent a non-targetlike lexical representation (photograph) rather than anything intonational *per se*, thus weakening the evidence for the kind of transfer being examined in Research Question 3. However, these 2 cases represent only a minority of the 14 total sentences under examination. For the remainder of these sentences, this problem does not arise since the Japanese accent location in the relevant word either matches up with English (e.g. ジン 'dinner') or falls on an epenthetic vowel inside the targetlike stressed syllable (e.g. メルマン 'mailman'). As such, for sentences containing such words, even under a loanword-influence account, there would be no reason to expect a Japanese EFL learner to begin the boundary rise from the final syllable, e.g. producing 'dinner'. As such, any distortion due to the above two sentences should be relatively minimal and therefore not compromise the validity of results for Research Question 3 as a whole.

5.3 Rating data

Recall from Chapter 4 (§4.2) that in order to assemble the results for all three research questions in terms of a broader hierarchy (i.e. which kinds of transfer occur more/less often), all three research questions share a common structure: "how often do Japanese EFL learners do ___?". Since the answer to this question may very well be different for beginning and advanced learners (since excessive transfer is a marker of one's being the former), the analysis needs to

---

8 Note also that the loanword for 'curry' (karē or, less commonly, karī) is unaccented in Japanese. This could potentially lead to Japanese EFL learners avoiding putting a pitch accent on 'curry' in the relevant sentence (*Does creole cooking use curry?*).
take proficiency into account. The fact that the corpus creators targeted a wide range of proficiencies in the design of the corpus (as mentioned above in §5.1.1) underscores the importance of doing so. With such information, it would be possible to surmise a possible developmental trajectory for the selected domains of L2 intonation knowledge (in apparent/pseudo-time).

Using such information, one approach would be to restrict the analysis to one specific proficiency level. For example, it would be possible to analyze only high-proficiency learners, which would have the advantage of factoring out the fluency and processing difficulties associated with lower-proficiency learners, hence the resulting data would be more directly indicative of intonation transfer specifically. However, such an approach would be effectively throwing away more than half of the available data, making the resulting picture incomplete (and perhaps even misleading). In particular, high-proficiency learners would potentially also have ceased not only having fluency and processing difficulties but also showing the transfer phenomena in question, making it difficult to draw any useful conclusions about the three research questions of interest. As such, in order to obtain a more complete picture, proficiency information will be used as a continuous regressor for the three intonation transfer 'outcomes' explored in the three research questions.

Unfortunately, while basic demographic information was collected from the learners (including residential history, parents' origins, standardized test scores, and out-of-classroom learning experience), none of this information is included with the public release of the corpus. The only information available in the corpus for gauging learners' proficiencies is the rating data that is available for a subset of the corpus, whereby native speakers of American English
provided several kinds of ratings of the learner productions, thus producing a multidimensional
gauge of overall phonological knowledge. Of course, this is only an indirect measure of the
learners' abilities in L2 phonology since (1) a rater may respond to irrelevant aspects of the
signal, and since (2) not all aspects of a learner's phonological knowledge will be reflected in
their ratings. This flaw notwithstanding, the rating data can nonetheless be useful as a proxy for
proficiency - and indeed incorporating some gauge of proficiency in the analysis is preferable to
including nothing at all. Furthermore, since this measure taps what native speakers thought about
learners' productions, it also has the advantage of ecological validity, since such information is
socially meaningful (in that a lack of accentedness is what one needs to demonstrate in order to
convince a native listener of one's phonological proficiency).

The rest of this section covers various issues relevant to these rating data. First, section
5.3.1 describes the overall structure of the rated portion of the corpus. Then, Section 5.3.2
discusses what information was available to raters during the process of performing the rating
tasks. Next, section 5.3.3 treats the topic of the instruction text shown to raters and the rating
scales they were asked to use. Section 5.3.4 then presents a global analysis of the relationship
between the various individual rating tasks. Finally, section 5.3.5 establishes why it is sound to
use rating data as a predictor for learners' intonation transfer outcomes.

5.3.1 Overall structure of rated portion of corpus

A subset of tokens in the corpus were rated by five phonetically-trained native speakers
of American English who had taught English in Japan. The final column ("Rated") in Table 5.1
above indicates the number of ratings available for each of the eight parts of the overall corpus.
As indicated there, only five of the eight parts of the corpus were rated, with 2 ratings obtained
for the words and either 4 or 5 ratings obtained for the sentences. The rating of these five parts took place in five separate tasks, administered through a web browser in sequence. For each of the five tasks, the raters were asked to make judgments about one specific aspect of the learners' pronunciation based on where the relevant items fall within the corpus design. Specifically, they were asked to rate segmentals for the phonemically balanced words and sentences, rate stress for the words varying in stress, rate intonation for the sentences varying in intonation, and rate rhythm for the sentences varying in rhythm.

Overall, 9484 (13.5%) of the 60,755 total tokens produced by the Japanese EFL learners are rated in this way. Ratings are available for the vast majority (190/202 = 94.1%) of the learners. The number of ratings is not balanced across items. Several of the phonemically balanced sentences were only rated once, whereas many of the word-stress items were rated as many as 26 times each. However, the number of ratings is relatively balanced across learners. Nearly every learner received 50 ratings (14.2%-14.6% of all tokens produced by each learner). These 50 ratings per learner can be broken down as follows:

---

9 The sole exception is learner "JE/TOK/F02", who received 34 ratings (9.8% of their data) total. More specifically, this learner was rated on 4 tokens for word segmentals rather than 20.
<table>
<thead>
<tr>
<th>Rating task</th>
<th>Corpus component</th>
<th>Tokens per learner</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word stress</td>
<td>Word stress</td>
<td>10</td>
</tr>
<tr>
<td>Word segmentals</td>
<td>Word segmentals</td>
<td>20</td>
</tr>
<tr>
<td>Sentence rhythm</td>
<td>Sentence rhythm</td>
<td>5</td>
</tr>
<tr>
<td>Sentence intonation</td>
<td>Sentence intonation</td>
<td>5</td>
</tr>
<tr>
<td>Sentence segmentals</td>
<td>Sentence segmentals</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Sentence rhythm and</td>
<td></td>
</tr>
<tr>
<td></td>
<td>sentence intonation</td>
<td></td>
</tr>
</tbody>
</table>

\[ \text{Total:} \ 50 \]

*Table 5.7: Breakdown of the 50 tokens rated per learner in terms of the five rating tasks and what component of the corpus the materials came from*

The above table makes it clear how there was mostly a one-to-one correspondence between what aspects of the corpus were rated for what. The only exception is the 'Rating sentence segmentals' task, where only half of the evaluated tokens came from the 'Sentence, Segmentals' materials, with the other half coming from some combination of the 'Sentence, Rhythm' and 'Sentence, Intonation' materials.\(^{10}\)

### 5.3.2 Information available during the rating process

For all five tasks, raters could see the relevant word/sentence in normal English orthography (hence it was always clear what the learners were trying to produce). In addition, for all three prosody-related tasks, the following visual information was also available. The effect of such visual information was presumably to constrain raters to base their ratings more closely on variance from these 'norms'.

---

\(^{10}\) The exact split between the R(hythm) vs. I(ntonation) materials varied learner to learner as follows: 5 R & 0 I = 107 learners, 4 R & 1 I = 13 learners, 3 R & 2 I = 19 learners, 2 R & 3 I = 16 learners, 1 R & 4 I = 10 learners, 0 R & 5 I = 25 learners.
(5)  

a. **Sentence, Rhythm**: Rhythm symbols of the sort shown in the second row of (1)

b. **Sentence, Intonation**: Up or down arrows indicating the direction of pitch movement at one or more points in the sentence

c. **Word, Stress**: TIMIT phoneme codes like those shown in the third row of (1), minus the stress-marking number suffixes. Instead, stress levels were indicated with color coding and font size.

In addition to this visual information, to clarify the meanings of the rhythm/intonation symbols, for the two sentence prosody rating tasks, i.e. (5a) and (5b), an audio recording was also available of the same sentences produced by a "model speaker" (a female native speaker of American English in her 50s from South Dakota). Since the audio file was placed at the top of the page for each sentence (above the area for rating learner productions), the raters were implicitly encouraged to listen to the model speaker first and establish a baseline before listening to the learners' productions. Thus, for these two tasks, raters were essentially asked to make an explicit comparison between the learners and the model speaker.

When making their responses, raters could listen to each learner token (as well as the model speaker) as many times as they wished. Moreover, due to the nature of the web-browser interface, raters could revisit and modify previously-given answers at any time. Thus, it is likely some responses for some raters were based on direct comparisons between learner tokens. While this would mean that each 'trial' is not statistically independent, since there is no record of rater responses at this level of granularity, it is unfortunately impossible to know how pervasive such cross-trial influence was.

**5.3.3 Instruction text and rating scale**

Due to the different focuses of the five tasks, the exact instructions differed from task to task. The complete instructions for all five tasks are provided below.
(6) a. **Sentence, Segmentals:**

Listen to each of the following sentences and evaluate the accuracy of segmental features in the connected speech, including linking, reduction, and allophonic variants. Evaluation should be as objective as possible and score the accuracy based upon a five-point scale.

b. **Sentence, Rhythm:**

Listen to each of the following sentences and evaluate the accuracy of English rhythm in the connected speech. Notice:

A. Sentences with rhythm patterns which were referred to when the students read the sentences are on the display. Refer to the patterns during your evaluation.
B. A model speaker's utterance is prepared for each of the sentences. Listen to each sentence of the model utterances before evaluating the learners' pronunciations of the sentence.
C. Evaluation should be done on the assumption that the correct rhythm is acoustically realized on the model utterance.

c. **Sentence, Intonation:**

Listen to each of the following sentences and evaluate the accuracy of intonation patterns in the corrected speech. Notice:

A. Sentences with intonation patterns which were referred to when the students read the sentences are in a PDF file. Open and display the PDF file with acrobat reader and refer to the patterns during your evaluation.
B. A model speaker's utterance is prepared for each of the sentences. Listen to each sentence of the model utterances before evaluating the learners' pronunciations of the sentence.
C. Evaluation should be done on the assumption that the correct intonation pattern is acoustically realized on the model utterance.

d. **Word, Segmentals:**

Listen to each of the following words and evaluate the accuracy of segmental features, including reduction, and allophonic variants. Evaluation should be as objective as possible and score the accuracy based upon a five-point scale.

e. **Word, Stress:**

Listen to each of the following words and evaluate the accuracy of lexical accent produced in the utterances in both terms of where the stress is and how the stress is. Japanese students tend to realize lexical accent only by pitch movement. Notice:

1. Words with their lexical accent labels which were referred to when the students read the words are on the display. RED, GREEN, and BLUE vowels mean the primary stress, the secondary stress, and no stress on the syllables including the vowels.
2. You will hear word phrases, not words, in No. 68 to 77. In these cases, RED, GREEN, and BLUE labels are assigned to words not to vowels. Namely, word-level accent label assignment. Of course, the students read the phrases by looking at these word-level labels. You have to judge the accuracy of prosodic aspects of the utterances as phrases.
Raters were asked to give their responses on a Likert scale ranging from 1 to 5. Under the relevant set of instructions for each task, the following definition was given for the rating scale to be used. (The wording of the scale was identical for all five rating tasks with one exception - for the 'Words, Segmentals' rating task, all instances of 'sentences' in the scale labels were replaced with 'words'.)

(7) 1. **Very poor** (inaccurate in pronouncing sentences, and apt to be misunderstood)
2. **Poor** (inaccurate in pronouncing sentences, and considerable practice needed)
3. **Fair** (fair in pronouncing sentences, and in intelligibility)
4. **Good** (accurate in pronouncing sentences, but some practice needed)
5. **Excellent** (good in pronouncing sentences, and very good in intelligibility, near-native speaker level)

In addition to the added 'bonus' trait *near-nativelikeness* at the highest level, the five levels in this scale can each be broken down into two halves as in Table 5.8. Indeed, raters may have used this structure (implicitly or explicitly) when performing the tasks. Note the model speaker (whose audio files were available to raters for the sentence rhythm and sentence intonation tasks) are not referenced anywhere in these clines.

<table>
<thead>
<tr>
<th>Level</th>
<th>Cline A</th>
<th>Cline B</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>inaccurate in pronouncing sentences</td>
<td>apt to be misunderstood</td>
</tr>
<tr>
<td>2</td>
<td>inaccurate in pronouncing sentences</td>
<td>considerable practice needed</td>
</tr>
<tr>
<td>3</td>
<td>fair in pronouncing sentences</td>
<td>fair in intelligibility</td>
</tr>
<tr>
<td>4</td>
<td>accurate in pronouncing sentences</td>
<td>some practice needed</td>
</tr>
<tr>
<td>5</td>
<td>good in pronouncing sentences</td>
<td>very good in intelligibility</td>
</tr>
</tbody>
</table>

*Table 5.8: Two clines underlying the scale used for the five rating tasks*
Cline A contains the parallel structure "___ in pronouncing sentences/words", with the remaining word creating the scale inaccurate < inaccurate < fair < accurate < good.\textsuperscript{11} These words describe both accuracy ("(in)accurate") and skill ("fair/good"). The latter presumably asks raters to assess the learner's level, e.g. how they would be classified on a standardized test. Cline B, on the other hand, describes both intelligibility ("fair/very good in intelligibility", "apt to be misunderstood") and preparation ("some/considerable practice needed"). The latter presumably asks raters to assess how much (self-)instruction a given learner still needs.\textsuperscript{12} To sum up, each of these clines taps two theoretically independent traits: accuracy and skill for Cline A and intelligibility and preparation for Cline B. Thus, between the two clines, at least four separate traits are collapsed into a single scale.

The facts sketched out above make the scale problematic for two reasons. First, it is not clear which cline raters deemed more important in assigning scores. This may have led different raters weighting the two clines differently, thus creating inconsistencies in the rating process (Schaefer 2008, Zhang & Elder 2011). Second, even if the two clines were weighted the same across all raters, it is still hard to interpret the rating data since both clines conflated multiple traits. In particular, it is unclear how levels of any one given trait correspond to levels of the scale (e.g., in Cline A, what a rating of 4 would mean for skill or a rating of 5 for accuracy). Considering how differences between raters' behavior can be rather extreme when different traits

\textsuperscript{11} It has been shown that parallel structures such as this simplify the task for raters, who latch on to them and heavily weight the key words therein during the scoring process (Hudson 2005).

\textsuperscript{12} Since the raters had all taught English in Japan, this 'preparation' construct may have caused raters to conceive of the speech samples as coming from students to be evaluated (pedagogically) rather than language-users to be assessed (analytically).
are rated separately, when all traits are put into a single scale (as here), it would be premature to assume the raters are all using that scale in the same way (Hudson 2005, Schaefer 2008).

While these shortcomings are unfortunate, they were in fact necessary due to practical constraints. The rated portion of the corpus comprises 27,439 judgments spread (unevenly) across five raters. With so many tokens to evaluate, it would have been impractical to have each rater provide multiple ratings of the same token, each with a different scale for a different trait.

### 5.3.4 Global analysis of relationship between rating tasks

One piece of information that is useful to have about the five different rating tasks is whether certain tasks are more 'skewed' towards ratings of 5 than others (e.g. ratings of rhythm as opposed to intonation). The different skews across the five different rating tasks are illustrated in the following figure:

**Figure 5.1: Global skews in the five rating tasks (raw ordinal data)**

This figure shows the overall frequency of each possible response (ratings of 1 through 5) for all five tasks, with all learners, raters, and items pooled together. The five panels in the figure have
been arranged from left to right in order of most to least skew towards ratings of 5. Note that
responses of 4 and 5 are almost equal in the leftmost panel whereas the rightmost panel is
centered at 3 with a slight skew towards 2.

Since the data are inherently ordinal, it is technically statistically unsound to treat the
responses as metric. However, it is interesting to note that the same pattern emerges if each
learner's overall average rating is calculated:

![Figure 5.2: Skews in the five rating tasks (Averages for each speaker)](image)

This figure is a histogram of speaker-by-speaker average ratings, with the same ordering
of panels from left to right. The overall mean average across all 190 speakers for each task is
indicated with a thick black dashed horizontal line. Note that this line gradually goes down from
the leftmost panel towards the rightmost panel. If the admonition against treating ordinal data as
metric is again ignored and the distributions in each neighboring pair of plots in this figure is
compared with a t-test, all differences are highly significant except the comparison between
Word Segmentals and Sentence Rhythm. These results can be summarized in the following
hierarchy:
Table 5.9: Overall hierarchy of skews in the five rating tasks

This hierarchy is broken into four tiers from left to right. Any horizontal comparison (i.e. one that crosses one or more cells from left to right) is significant, whereas any vertical comparison is not. One interesting pattern emerging from these results is that words (as a whole) are ranked higher than sentences. This is perhaps tied to the amount of linguistic material raters had to judge; with a sentence, there are more possible areas where a learner's non-nativeness could surface (and therefore could be deducted). Secondly, within each of these two clusters (for words/sentences), segmentals are always rated lowest. This could be either because Japanese EFL learners struggle with segmentals the most or else because the native-speaker raters' intuitions about 'correct' prosody are more flexible.

In addition to the inherent skews within each individual rating task, another useful piece of information about the rating data is how the results for the different tasks are related to each other. Toward this end, the following table shows the correlations between all pairwise combinations of rating tasks.\(^\text{13}\) Once again, the data points going into each correlation represent speaker-by-speaker averages, and once again the ordinal data are treated as metric.

\(^{13}\) Note that the correlations reported in this table and the rest of this section represent \(r\) values, not \(r^2\) values. Of course, the latter can easily be calculated from the former.
Table 5.10: Correlations between speaker-by-speaker averages for the five rating tasks

To make it easier to grasp the big picture that emerges from this table, it can be helpful to visualize these correlations in terms of a spatial representation, as in the following figure.

Figure 5.3: Multi-Dimensional Scaling visualization of correlations between the five rating tasks

Since all correlations in Table 5.10 are positive, by calculating $1 - \text{(correlation value)}$, they can be mapped onto a scale from 0 (smallest 'distance' between tasks, i.e. highest correlation) to 1 (largest distance between tasks). This can then be fed into a standard Multi-Dimensional Scaling (MDS) analysis. The output (or 'solution') of this analysis is what
determines the location of the text labels (the rating task names) in Figure 5.3. All logically possible pairwise combinations of tasks are then connected with grey lines, and the corresponding correlations (rounded to 2 digits) are placed at the midpoints of these lines. The only ones left off of the plot are long-distance connections that would visually overlap with others and clutter up the visualization. Here, these are SentenceRhythm-WordSegmentals (correlation of 0.393) and SentenceIntonation-WordStress (0.339).

Interestingly, the two dimensions in this MDS solution map transparently onto readily recognizable constructs. The horizontal dimension indexes the word (far right) vs. sentence (left) distinction. The vertical dimension, on the other hand, indexes the segmentals (top) vs. prosody (intonation/rhythm/stress, bottom) distinction. Thus, in a way, this MDS solution mirrors the two-by-two structure of the corpus as a whole (and the rating tasks thereof).

This visualization of the correlation table highlights several facts. All of the correlations weaker than 0.5 (including the two excluded from the plot) are (1) between one word rating task and one sentence rating task and (2) tapping two different domains (e.g. segmentals vs. intonation). This means that, in general, a learner having a high average rating for a word task does not guarantee that they will have a high average rating for a sentence task in an unrelated domain, and vice versa. This is a logical result considering that the tasks are maximally different from each other. The highest correlations of all are either two tasks on the same domain (e.g. Word Segmentals and Sentence Segmentals at 0.64) or two tasks both involving sentences (e.g. Sentence Segmentals and Sentence Rhythm at 0.68). Finally, on a more global level, all correlations are modest to strong (in the range of 0.339 to 0.684) but not extremely high (i.e.
above 0.9). This suggests that the raters were most likely tapping separate constructs when rating the different tasks, as is desirable.

5.3.5 Using rating data as a predictor

In the following chapter, these various rating dimensions will be used to predict learners' performance as it pertains to the present study's research questions. Since foreign accent is partially driven by non-native intonation, it may appear circular to use these ratings to predict the learners' intonational performance. However, this risk of circularity is expected to be minimal for two reasons. First, as just mentioned, these exact sentences were rated only on segmentals, not intonation. The fact that the intonation ratings come from a different set of sentences entirely helps to break the circularity. Second, an examination of Table 5.2 makes it clear that the intonation component of ERJ was not designed to test any of the three specific phenomena under examination in the present study. Rather, most of them deal with the broader issues of what tunes are conventionally associated with what kinds of texts. Given this design of the materials, the three specific phenomena investigated in the present study's research questions should have only exerted minimal influence, even in the 'sentence, intonation' rating task. Thus, despite the partial overlap in domain, the risk of circularity should be low enough not to invalidate the conclusions of the present study.

Very few of the exact sentences selected for the research questions (as listed in §5.2) were evaluated in the rated part of the corpus. For example, of the many hundred tokens available for Research Question 3, only 21 were rated (1 token for each of 17 learners plus 2 tokens for each of 2 additional learners). Since the selected sentences come from the [Sentence, Segmentals] portion of the corpus, all such ratings are for sentence segmentals (not anything
prosodic). With ratings only available for such a small subset of the selected sentences, it would be of little use to restrict the analysis to just those sentences and analyze them separately. As such, the approach taken here is to collapse all ratings at the level of the individual learner for purposes of analysis. More specifically, for each of the five rating tasks (sentence intonation, word segmentals, etc.), all ratings that each individual learner received are averaged into a single mean score. This means that each individual learner has five scores - one for each of the five rating tasks.

Unfortunately, the selection of the sentences for the three research questions did not control for list. Information on the distribution of the selected sentences into the lists is presented below. (This information can also be gleaned from the "List ID" column in the tables of selected sentences in §5.2.1 through §5.2.3.)

<table>
<thead>
<tr>
<th>List</th>
<th>Research Question 1</th>
<th>Research Question 2</th>
<th>Research Question 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>14</strong></td>
<td><strong>14</strong></td>
<td><strong>14</strong></td>
</tr>
</tbody>
</table>

*Table 5.11: Number of sentences in each list selected for each of the three research questions*

The most notable imbalance occurs for Research Question 1, where over half of the sentences come from List 3 and none come from Lists 4 through 8. This creates an imbalance in sampling, as the learners who read list 3 have over twice as many data points represented in the
sample compared to those who read Lists 1 and 2. Of course, it would have been ideal to ensure that the sentences were adequately structured for looking at rates within a single learner. However, this was not a criterion for selecting sentences from the ERJ for the purposes of the present dissertation, as it lay outside the scope of its immediate goals. Moreover, limitations in the materials in the ERJ would have made this impossible.

To rectify this problem, the approach taken here is to group learners into three categories - High, Mid, and Low - for each of the five rating criteria. The cutoff between the three groups as the 33.33% and 66.67% quantile of the distribution for a given criterion across all 190 learners. Any learners falling exactly on a cutoff are treated as Mid. The exact cutoffs for each of the rating tasks are provided in the following table.

<table>
<thead>
<tr>
<th>Rating criterion</th>
<th>Low-Mid cutoff</th>
<th>Mid-High cutoff</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sentence Intonation</td>
<td>2.9</td>
<td>3.35</td>
</tr>
<tr>
<td>Sentence Rhythm</td>
<td>3.15</td>
<td>3.7</td>
</tr>
<tr>
<td>Sentence Segmentals</td>
<td>2.66</td>
<td>3.14</td>
</tr>
<tr>
<td>Word Accent</td>
<td>3.65</td>
<td>4.05</td>
</tr>
<tr>
<td>Word Segmentals</td>
<td>3.2</td>
<td>3.625</td>
</tr>
</tbody>
</table>

*Table 5.12: Cutoffs used to divide learners into Low/Mid/High 'proficiency' groups*

Since these cutoffs are defined separately for each of the rating criteria, this approach takes into account the different skews inherent to the different rating tasks discussed in the previous section (§5.3.4). Moreover, since this is defined based on the entire sample of 190 learners (irrespective of the specific sentences selected for a given research question), this avoids
the problem of imbalance (documented in Table 5.11) across the sentences selected for the three research questions.14

Whereas this section (§5.3) documented how the ERJ's rating data was handled, the following section (§5.4) describes how data coming from the soundfiles that make up the ERJ were handled.

5.4 Data processing workflow

This section details how the data selected for the three research questions was processed. The data processing took place in three stages. First, tokens with deviations from the script (e.g. with a word missing) needed to be identified and, if necessary, discarded (§5.4.1). Second, the remaining utterances were then 'segmented', i.e. annotated for the location of the boundaries between segments/syllables (§5.4.2). Finally, using the method described in Chapter 3, the F0 track extracted from each token was stylized (§5.4.3).15

5.4.1 Identifying segmental deviations

While great care was taken in the data collection process to make the data as homogeneous and error-free as possible, learners and natives alike occasionally diverged from the script in unpredictable ways (e.g. beyond what can be expected from normal epenthesis processes), which occasionally could potentially impact the analysis. As such, each soundfile was first listened to individually and all segmental deviations from the script were systematically

---

14 Since only 190 of the 202 total learners in the ERJ have rating data available, a small fraction of the learners analyzed for a given research question have no rating data available. Such learners are simply excluded from the analysis.

15 The core annotation data analyzed in the present dissertation (F0 ranges, segmentations, and stylizations) are freely available in the IUScholarWorks repository at http://dx.doi.org/10.5967/K86Q1V51.
cataloged. Then, the full list of all such deviations was inspected, and it was determined which of these would have an impact on the conclusions regarding the three research questions under investigation. In most cases, the divergences affected the utterances in ways irrelevant for the research questions at hand (e.g. from altering a portion of the utterance outside of the target region). However, a small minority of tokens could have possibly compromised the results and were therefore excluded. Such exclusions were based on segmental deviations only, as reference to prosodic criteria would have run the risk of circularity. (For example, in some sense, a 'wrong stress' is precisely what is of interest for Research Question 3.) The exact details of what was excluded and why are presented separately for each research question below.

For Research Question 1, what is of crucial interest is the number of syllables before the primary stress. Thus, any tokens with deviations from the script that did not affect the number of syllables in this region (e.g. swapping out or adding a single consonant) were included. Examples of this type are amb[ar]dextr[ks], continen[f]al, Vietnami[n], Vietnam[ar]z, and artifi[k]al. The number of syllables after the primary stress was allowed to be different, e.g. artificial[i] (as an adverb) or ambidextr[i.ә]s. Only eight tokens were excluded, which fell into two categories. Three tokens had a syllable deleted in the crucial pre-primary-stress region (.__.ca.de.mic, en.__.clo.pe.di.as, __.cy.clo.pe.di.as). In five other tokens, the primary-stressed syllable itself was deleted (ca.ta.____.pic, am.bi.____.trous, em.ploy.____ (x2), Vi.et.nam_s). All eight of the excluded tokens were from learners, thus dropping the grand total of learner tokens for Research Question 1 from 327 to 319. (The total number of tokens for native speakers remained unchanged at 154.)
For Research Question 2, what is of crucial interest is the number and type of tonally-marked prosodic boundaries in the utterance. Many tokens contained instances where articles ('the' or 'a') were deleted, inserted, or substituted for each other. Since both of these words are only monosyllabic function words, none of these kinds of change were expected to significantly affect the number of boundaries. Only eight tokens were excluded. All of these had disfluencies of one kind or another, such as a false-start plus repetition, glottal stop insertion, or inserting long pauses in the middle of a word. Disfluencies of this sort generally occurred when the speakers caught themselves from accidentally reading the text incorrectly (e.g. when processing resources cannot keep up with the reading task). Since disfluencies artificially inflate the number of prosodic breaks in irrelevant ways, in the effort of being conservative, they were excluded from the analysis. Half of these came from learners and the other half from natives, thus dropping the overall counts of learners+natives from 363+154 to 359+150.

For Research Question 3, the critical region is the string of unstressed syllables after the nuclear stress in the final word. Any segmental deviation that left this region unchanged was included, either because the deviation affected some other word earlier in the sentence (e.g. 'this photograph' instead of 'the photograph') or because it only added or substituted consonants in the final word (e.g. 'biddant' or 'biddings' instead of 'bidding'). Only five tokens needed to be excluded for changing the number of syllables in the final word: (1) employment → empoliments, (2) removal → remove, and (3) forces → force (x3). All five of the excluded tokens were from learners, thus dropping the grand total of learner tokens for Research Question 1 from 347 to 342. (The total number of tokens for native speakers remained unchanged at 154.)
5.4.2 Segmentation

Next, each file in the dataset was segmented into labeled intervals representing the segmental makeup of the utterance. This step was performed with reference to both a visual display of the spectrogram as well listening to the audio. A module specially programmed in R for this purpose was used for this task. The module drew upon the R's 'audio' library and the Spectrogram() function at https://github.com/usagi5886/dsp/.

Silences were explicitly marked as "<SIL>". Such silences usually had one of two sources: either (1) pausing mid-utterance or (2) a closure for a stop or affricate consonant at a boundary between two intervals. Both kinds of silences were marked, regardless of the silence's duration, because, for example, before a [p], the distinction between a closure of 'normal' duration and a full-fledged pause is gradient and continuous, rather than categorical, in nature.

Not surprisingly for this learner population, epenthesis (and to a lesser extent, deletion) processes were ubiquitous in the data. Adding extra segmentation boundaries in an unpredictable way on a token-by-token basis would have significantly complicated the analysis. Instead, the approach taken was to group the extra segment along with the adjacent vowel with which it would be syllabified in Japanese. Thus, if an extra [u] were inserted at the end of 'catastrophic', since this would create the Japanese syllable /ku/, the [k] and vocalic portion were both grouped into the segmentation interval for [k]. This effectively means that the fully targetlike production for a given word was used as a template, or frame of reference, for parsing the learner's production. As such, the segmentations of learner productions are not intended to be a representation of the phonetic form actually produced by the learner. Rather, it just partials out their production in terms of the targets, i.e. what they were presumably trying to produce, thus
establishing a common frame of reference from which all tokens could be compared on equal
grounds. In addition to accounting for learners' segmental deviations, this methodological choice
also helps factor out variant pronunciations on the part of native speakers (e.g. reducing
/ˈvaɪ.ə.lɛns/ to [ˈvaɪ.ә.lɛns] or even disyllabic [ˈvaɪ.lәns]).

The exact domain to be segmented differed for each research question. For Research
Question 1, only the first word was of interest, hence that was the only portion of the utterance
that was segmented. To give maximum granularity, the segmentation was performed at the level
of *visually distinct sets of phones*. In cases where a consistent division between phones could not
be expected (e.g. for the different qualities inside a series of vowel sounds), they were combined
into the same interval. Thus, for example, in 'scientific', finer distinctions within the vocalic
sequence [aɪә] could not be readily distinguished and therefore were not marked.\textsuperscript{16}

The following table contains the full list of intervals used for this research question.\textsuperscript{17}

\textsuperscript{16} Grouping phones together in this way has the disadvantage of requiring some arbitrary cutoff, blind to the fine-
gained time-varying spectral characteristics of the signal, to be applied later in the analysis in order to segment
individual phones (e.g., in 'scientific', declaring the first 2/3 of the interval to be [aɪ] and the last 1/3 to be [ә]).
However, the advantage of this approach is that the annotator is not forced to make questionable decisions in the
segmentation stage, which would make the segmentation a less accurate representation of the original signal and the
entire analysis less replicable.

\textsuperscript{17} Since the target words are utterance-initial, silences produced in the middle of a target word (e.g. *cata...strophic*)
were exceptionally rare (only N=2). This suggests that problems tied to fluency or processing generally did not
affect the data for this research question.
Table 5.13: Segmentation intervals marked for Research Question 1

For Research Question 2, the entire sentence was segmented into individual syllables. Consonants are generally syllabified into onsets wherever possible (e.g. heating) except where that would cross a word boundary (e.g. lack|of rather than la|-ckof). The following are the exact intervals used for each sentence:

<table>
<thead>
<tr>
<th>#</th>
<th>Word</th>
<th>Individual segments</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>catastrophic</td>
<td>kætəstərafɪk</td>
</tr>
<tr>
<td>2</td>
<td>ambidextrous</td>
<td>æmbɪdɛksˈtraʊs</td>
</tr>
<tr>
<td>3</td>
<td>academic</td>
<td>ækədɛmɪk</td>
</tr>
<tr>
<td>4</td>
<td>continental</td>
<td>kæntəˈnɛntl</td>
</tr>
<tr>
<td>5</td>
<td>cooperation</td>
<td>kəʊəpərˈeɪʃn</td>
</tr>
<tr>
<td>6</td>
<td>masquerade</td>
<td>mæskəˈeɪd</td>
</tr>
<tr>
<td>7</td>
<td>encyclopedias</td>
<td>ɛnˈsaɪkləpɪdiəz</td>
</tr>
<tr>
<td>8</td>
<td>artificial</td>
<td>əˈtɜːfɪˈl</td>
</tr>
<tr>
<td>9</td>
<td>pizzerias</td>
<td>pɪtəˈriəz</td>
</tr>
<tr>
<td>10</td>
<td>scientific</td>
<td>səˈsentɪfɪk</td>
</tr>
<tr>
<td>11</td>
<td>agricultural</td>
<td>ægəˈrɑltʃərəl</td>
</tr>
<tr>
<td>12</td>
<td>Vietnamese</td>
<td>vɪɛtnəˈmɪz</td>
</tr>
<tr>
<td>13</td>
<td>curiosity</td>
<td>kjuˈəsətɪ</td>
</tr>
<tr>
<td>14</td>
<td>employee</td>
<td>ɪmˈpləʊi</td>
</tr>
</tbody>
</table>
Table 5.14: Segmentation intervals marked for Research Question 2

Since the crucial region for Research Question 3 is the utterance-final word, only that word was segmented. As with Research Question 1, the segmentation was into individual visually distinct phones. The following table displays the exact intervals used:
## Table 5.15: Segmentation intervals marked for Research Question 3

The total number of boundaries (including silences) thus marked was 4,559 (Research Question 1), 11,224 (Research Question 2), and 3,295 (Research Question 3). The number is significantly larger for Research Question 2 simply because every syllable was marked, and the selected sentences for this research question were among the longest in this entire portion of the ERJ. With all three research questions combined, a grand total of just over 19 thousand (N=19,078) boundaries were manually marked for the present dissertation.

### 5.4.3 Stylization

Next, Pitch object textfiles were created for each individual utterance using PraatR (Albin 2014) as described in Chapter 3 (§3.1). In order to minimize F0 tracking errors (e.g. 'pitch halving' or 'pitch doubling') as much as possible, the F0 range was manually fine-tuned for each individual file. Even in cases where only one specific word inside the sentence was of crucial
interest, the F0 range was set so that the F0 values across the entire sentence would be accurately represented.

With the F0 information thus extracted, the F0 track inside each file could then be stylized as described in Chapter 3, thus discretizing the noisy raw signal into a series of F0 targets ('vertices') and the nonlinear F0 movements between them ('transitions'). The stylization process (in particular, the process of selecting vertices) was done entirely manually. To do so, rich visualizations like those described in Chapter 3 (§3.1.3) were displayed, containing the waveform, F0 track, and spectrogram for the utterance. In addition, during the entire stylization process, vertical lines cutting across all three panels showed the segmentation of the utterance into the intervals listed in the previous section (§5.4.2). The vertices were then manually selected using the identify() function in R, which prompts the user to select specific points from a plot. After the vertices for a given contour were manually selected, the shapes of the transitions between them were determined automatically. As mentioned in Chapter 3 (§3.4.4), this involved determining the gradience and threshold parameters that best describe the transition shape using a standard non-linear optimization algorithm.

The exact domain that was stylized depended on the research question. Research Question 1 (about phrasal H-) focuses on what happens intonationally across the first word in the utterance. As such, only the F0 track across this one word in each sentence was stylized. Only the portion of the signal across this first word was displayed, and all vertices placed inside this time range. Research Question 2 (about frequency of L% usage) concerns global F0 patterns across the entire sentence. Consequently, the entirety of each relevant sentence was stylized for this research question.
For Research Question 3 (about the alignment of L% in boundary rises), it is important to see the broader pattern of the contour, even beyond the sentence-final word, in order to detect cases where the nuclear pitch accent occurs early in the sentence. As such, unlike Research Question 1, the entire utterance was displayed during the stylization process, and vertices could be placed anywhere in the utterance. In order to distinguish between single vs. double rises, the final two F0 transitions of the utterance were stylized - the final rise itself plus the immediately preceding transition. Put another way, the analysis was restricted to the final three turning points ('vertices') in the F0 contour, regardless of where they occurred.

All in all, the total number of vertices thus marked was 1,951 (Research Question 1), 7,394 (Research Question 2), and 1,503 (Research Question 3). As was seen to be the case above for segmentation (§5.4.2), the number is significantly larger for Research Question 2 because the entire sentence was stylized (rather than a single word) and the sentences for this research question were among the longest. With all three research questions combined, a grand total of just over ten thousand (N=10,848) vertices were manually marked for the present dissertation.

In the present context, the F0 track serves as a record of what some speaker 'did' in a given utterance. The stylization process requires weighing various hypotheses about what the underlying gestural structure behind that utterance could be. The final analysis captured in the stylization is the hypothesis found to be the most persuasive. This generally involved listening carefully to get a clear idea of what sort of information should be stored in the stylized representation, finding where each stretch of curvature begins and stops, and choosing vertices so the desired information is represented correctly. In this process, the nonlinear curve shape is crucial since it helps make it clear whether the selected vertices are a good match to the data (i.e.
whether a given hypotheses is a reasonable one). When the line for a given transition is a poor fit to the raw data (i.e. when the model is bad), it is readily apparent. This is a significant advantage over intonation transcription (e.g. ToBI or RaP), where there is no such feedback loop to keep one's modeling in check.

It often took several tries to get the selected vertices to actually constitute an accurate model of the F0 contour. As such, the 'undo' feature built into the stylization module was used very frequently. The stylization process continued until the model successfully captured all of the desired information in the rich visualization. At that point, everything not captured in the model (i.e. the residuals off of the transition lines) was deemed to be safe to exclude (e.g. if they were attributable to well-documented segmental perturbations for obstruents). The final stylization for each token was stored as a matrix in a textfile. (For the exact format of this data storage representation, refer back to the example in Chapter 3 (§3.5).)

The core decision rule used in the stylization process was to *err in ambiguous cases on the side of fewer targets*. This decision rule manifested itself in two different ways. First, generally speaking, long flat stretches were represented by a single vertex (at the endpoint of the flat stretch) rather than two (one at each end). This can be thought of as more directly representing the motoric commands to the larynx, since the beginning of the flat stretch is often analyzable as merely a (often vaguely-defined) region around which the majority of that one single F0 movement was achieved. Second, placing vertices at both sides of a pause was generally avoided. Typically, it was possible to analyze the F0 movement around that pause as involving only one F0 target before the pause, representing some boundary tone. In such cases, the F0 after the pause merely 'picked up where it left off' and transitioned towards the first vertex
thereafter (rather than beginning from some phonologically-specified value). The minority case where a second vertex was required after the pause involved cases of F0 range reset.

The only exception to this decision rule occurred in cases where it would be possible to analyze a stretch of a contour as either a single transition with an abrupt movement (e.g. suddenly jumping up 50 Hz), as indexed by a high value for the gradience parameter of that transition. In such cases, a second decision rule was invoked, whereby *transitions should only be moderately nonlinear*. Thus, in the process of creating stylization models, if a transition ever exhibited this kind of abrupt movement, it was treated as a red flag and an alternate model was pursued with an additional vertex in the middle (allowing the transitions to have lower gradience).

Since it is driven by holistic human judgment, the process of manually coding the data as just described is, by design, imperfect. While the procedures described above attempt to make the stylization process transparent and reliable, there is no guarantee that the coding reflects the learner's actual intended phonological intent (i.e. their motoric commands). However, this is true of almost any analysis that can be conceived. Moreover, the manual annotations are no less noisy than the raw data itself (and indeed the purpose of the annotations in the first place is to drastically reduce the amount of noise). Furthermore, as discussed at the end of §3.3.2 in Chapter 3, the 10 thousand data points of manual annotations produced here can serve as the gold standard for algorithms that automate the stylization process; hence, from a broader perspective, the flaws of this approach may be short-lived. Finally, as mentioned in §1.2, the stylization approach adopted here achieves the crucial separation between the description of what a learner
does with F0 and the interpretation as to why they might have done that (i.e. phonological hypotheses about tonal makeup) - something that cannot be said for transcription.

Moreover, using the intensity-weighted median absolute deviation (‘weighted MAD’) criterion discussed in §3.4.4, it is possible to provide an independent measure of the quality of the stylizations. If all 9421 transitions from all 1353 stylized contours are pooled together, the distribution in weighted MADs looks like the following.

![Distribution of weighted MADs](image)

*Figure 5.4: Distribution of weighted MADs across all three research questions in raw scale (left) and (natural) log scale (right)*

On the raw scale, the median is 0.2742075 and the range is 0 to 6.29173. (If this distribution is log transformed with the natural logarithm, the median is -1.29387 and the range is -Inf to 1.839236.) Taken together, this set of facts suggests that the distribution of weighted MADs is very heavily skewed towards 0. Since low values for weighted MADs indicate a good fit to the data (especially over high-intensity portions), this is precisely what should be the case if
the stylizations were an accurate reflection of the data. This empirical evidence further bolsters the theoretical arguments made above that the stylization process produces sufficiently high-quality data, at least for the purposes of the present dissertation.

With the data thus processed, any given token is represented in the data by a combination of one segmentation plus one stylization. These two pieces of information can then be integrated to conduct the various analyses required for obtaining answers to the research questions. The following chapter outlines the makeup of these analyses as well as the results thereby obtained.
Chapter 6: Results

This chapter discusses the empirical results of the present dissertation. The approach adopted here for getting the answers to the research questions is to write queries for the stylizations and segmentations. This involves pulling out all instances of some specified structure (the phenomenon of interest for each research question) and seeing how many are returned. Since the stylizations are essentially phonetically-explicit hypotheses about the gestural structure underlying speakers' F0 tracks (as mentioned in §5.4.3), this equates to analyzing global patterns and identifying common trends across all such hypotheses. By seeing how frequently certain patterns appear in learners (as opposed to native speakers), this makes it possible to get straightforward answers to the research questions.

In each case, the query run only taps a small set of piece of a much bigger, more complicated picture. For example, Research Question 3 boils everything down to a small set of measurements describing where the last two F0 transitions occur. While the tokens vary widely in the shape of the nuclear pitch accent - sometimes English L+H*, other times Japanese H*+L, etc. - such information is only indirectly relevant and therefore not captured in the analysis. The queries seek answers to very specific questions, collapsing across these various other surrounding factors that occur in the tokens.
Each of the three research questions is discussed separately in §6.1 through §6.3 below. For each, the discussion is broken into four smaller subsections. First, the structure of the query for the relevant research question is discussed (§6._.1). Then, the results are presented regarding the overall relative frequency of the transfer phenomenon (§6._.2). The next section (§6._.3) examines the extent to which the relative frequency can be statistically predicted by the rating data. Finally, the various analyses are synthesized in order to address the corresponding research question (§6._.4).

Before delving into the discussion of the individual research questions, it is worth pausing briefly to illustrate of how the raw data flowing into the analysis is structured. Consider the following token, representing the first word of the Research Question 1 sentence *Masquerade parties tax one's imagination*, as produced by an English native speaker.
The stylization in this plot contains four vertices: the first vertex (which has no preceding F0 transition) followed by a rise, a fall, and then another (somewhat smaller) rise. This information, along with the temporal alignment of each vertex relative to the segmentation boundaries, can be extracted out of this plot and summarized in an even higher-level representation as in the 'dot plot' visualization below. This dot plot succinctly summarizes the stylization and segmentation information for all tokens of the 'masquerade' sentence (spoken by both natives as well as learners).
Figure 6.2: Visualization of overall data structure of a single sentence (Research Question, 1 sentence 6, "Masquerade...")

Along the top of the dot plot, the IPA transcription of the target word ('masquerade') is provided, divided by dotted grey vertical lines into the corresponding segmentation intervals. Along the y axis, speakers are grouped into learners (top) and natives (bottom), divided by a
solid black line. Each individual token is represented as a 'row' of points in the plot in the relevant region. Within the two regions (for learners and natives), the order of these rows is arbitrary, reflecting only the speaker's ID number (in increasing order from top to bottom). The native stylization in Figure 6.1 corresponds to the very bottom row.

Positive F0 changes are indicated with grey unfilled upward-pointing triangles, and negative F0 changes are indicated with black filled downward-pointing triangles. Thus, note that in the bottom row, the second and fourth vertices (where there are F0 rises in the F0 track) are represented with grey upward-pointing triangles. The first vertex in each utterance (for which there is no syntagmatic change value) is indicated with a '×' symbol.

The location of individual points along the x axis is time-normalized in terms of percentages across the relevant interval. For example, if a [m] extended from 100 ms to 200 ms in a soundfile and a vertex occurred at 163 ms, then this would be placed at 63% across that interval. Thus, note how the temporal location of the vertices are in the bottom row of the dot plot match up with Figure 6.1. To emphasize the fact that the vertex locations are time-normalized, all intervals share a fixed width (of 1) in the figure. Note that, if the vertical dimension is ignored, a dot plot functions like a histogram describing the distribution of vertex locations.

Thus, the raw data that can be queried for each of the three research questions consists of a series of vertices (each with a syntagmatic F0 change value), time-normalized in terms of percent across the relevant interval in the segmentation. While various other pieces of information in the stylization (such as threshold, gradience, weighted MAD, or percent voiced frames) could in theory also be brought into the analysis, this is not done in the analyses below.
6.1 Research Question 1

The query described in the following section (§6.1.1) is based in part on the syllable affiliation of individual vertices. The default way of determining affiliations is simply whether the vertex in question falls within the set of intervals corresponding to the syllable in question. For example, the second vertex in Figure 6.1 occurs inside [æ], hence within the syllable 'mas-' of 'masquerade'. However, it was occasionally the case that vertices fell a few milliseconds into the syllable adjacent to where it is expected to occur. As a hypothetical example, consider the word 'aces', syllabified as [eɪ.sɪz] and segmented into the four intervals [eɪ], [s], [ɪ], and [z]. Since the [s] is a voiceless obstruent, it would normally not be assigned any vertices in the stylization process. (Note the lack of vertices over [s] and [k] in Figure 6.2 above.) However, due to minute measurement error (of various sources), a vertex that clearly should be affiliated with the [eɪ] may be found a few milliseconds into the [s] (e.g. 2% across the interval for [s]). Taking the exact millisecond boundary between [eɪ] and [s] in this case would therefore lead to a misclassification of that vertex's syllable affiliation (and therefore introduce noise into the dataset). In such cases, since the [s] is effectively an 'empty zone' in the set of stylizations (i.e. a long range of normalized time with no vertices), it is safe to simply adjust the syllable cutoff accordingly in an ad-hoc manner. In this example, the division point between the regions where vertices are associated with the first syllable [eɪ] vs. the second syllable [sɪz] could perhaps be placed at 50% across the duration of the [s].

In this way, cutoff locations were determined manually based on inspecting figures like Figure 6.2 for every sentence of Research Question 1. When a voiceless obstruent was adjacent to a syllable division, the cutoff was generally placed at 50% across the voiceless obstruent. In
the effort of making the analysis fully replicable, the following table lists all of the exact cutoffs used. A unit of 1 corresponds to one roughly segment-sized interval in the plot above (with the leftmost edge of the plot beginning at 0). Thus, for example, the first boundary in *catastrophic* occurs at 2.33, or 1/3 into the third interval, i.e. the [t] of [kæt].

<table>
<thead>
<tr>
<th>Word</th>
<th>Cutoffs</th>
<th>Word</th>
<th>Cutoffs</th>
</tr>
</thead>
<tbody>
<tr>
<td>catastrophic</td>
<td>2.33, 4.33, 8.33</td>
<td>ambidextrous</td>
<td>2.33, 4.00, 8.00</td>
</tr>
<tr>
<td>academic</td>
<td>1.33, 3.00, 5.25</td>
<td>continental</td>
<td>3.00, 4.50, 8.00</td>
</tr>
<tr>
<td>cooperation</td>
<td>1.50, 2.33, 3.75, 5.33</td>
<td>masquerade</td>
<td>3.00, 4.90</td>
</tr>
<tr>
<td>encyclopedias</td>
<td>2.00, 4.33, 7.33, 9.33, 10.50</td>
<td>artificial</td>
<td>1.33, 3.33, 5.33</td>
</tr>
<tr>
<td>pizzerias</td>
<td>3.00, 5.50, 6.55</td>
<td>scientific</td>
<td>1.50, 3.33, 5.33</td>
</tr>
<tr>
<td>agricultural</td>
<td>2.00, 4.33, 6.33, 8.50</td>
<td>Vietnamese</td>
<td>1.50, 3.00, 5.00</td>
</tr>
<tr>
<td>curiosity</td>
<td>1.33, 1.67, 2.33, 4.00</td>
<td>employee</td>
<td>2.33, 4.30</td>
</tr>
</tbody>
</table>

*Table 6.1: Syllable boundaries used for Research Question 1*

### 6.1.1 Structure of query

Before applying the query for Research Question 1, the stylized contours were first re-coded in terms of what F0 targets occurred inside each syllable. To do so, every transition was classified as either 'h' (high, i.e. rise), 'l' (low, i.e. fall), or 'e' (equal, i.e. flat). The criterion used for this classification was 50 cents (i.e. one half-semitone) syntagmatic F0 change from one vertex to the next.\(^1\) In other words, any transition of +50 cents or more was considered a rise (H), anything between +50 and -50 was considered flat, and anything under -50 cents was considered a fall (L). To give an idea of the magnitude of these changes, a change of +50 cents from 100 Hz

\(^1\) Given two F0 values \( \text{LeftF0} \) and \( \text{RightF0} \) in a transition from left to right, the value in cents is given by the formula \( \log_2(\frac{\text{RightF0}}{\text{LeftF0}}) \times 12 \times 100 \), where \( \log_2() \) refers to a logarithm with base 2. Note that a cent is 1/100 of a semitone, and a semitone is 1/12 of an octave. Crucially, there is no one fixed reference value in this calculation (as would be the case, e.g., if one were calculating semitones relative to 100 Hz or some other arbitrary value). Rather, the calculation of 'number cents change' is applied separately to every pair of vertices, such that any given vertex serves as the reference value for the one immediately following. For example, in the string of vertices \([A, B, C]\), A serves as the reference value for B, and B serves as the reference value for C.
would reach 102.93 Hz (a +2.93 Hz difference). In a higher area of the F0 space, the difference in Hertz is of even larger magnitude (thus mirroring the nonlinearities in human pitch perception), e.g. a change of -50 cents from 200 Hz would reach 194.3 Hz (a -5.7 Hz difference).

With the data thus re-coded, the queries for Research Question 1 could be run, seeking evidence for a phrasal H- category transferred from Japanese. The query designed to tap this attribute of learners' L2 English production can be thought of as having two 'stages'. In the first stage, a specific definition of what it means to be a phrasal H- was applied, and all tokens meeting this definition were retrieved. In the second stage, any superficially similar lookalikes within the L2 system were excluded. Any tokens that remained after these two stages were retained as potential candidates for bona fide transfer of the L1 phrasal H- category. In essence, this approach identifies any high F0 targets that are otherwise unexplainable from the perspective of the L2 system. Each of the two stages will now be explained in greater detail.

The first stage involved retrieving any tokens whose stylization met a specific definition of what it means to be a phrasal H-. The definition of this category has two components: scaling and alignment. The scaling (vertical) dimension is relatively straightforward - a phrasal H- needs to be at the endpoint of a syntagmatic rise (i.e. an 'h' in the recoding).² (The very first vertex of the utterance, for which no syntagmatic F0 change value is available, is by definition excluded.)

² Note that this definition only retrieves instances where a high target was explicitly marked over the appropriate syllable/mora in the stylization. It is also possible, however, for a Japanese phrasal H- to manifest itself as merely a gradual elbow in the transition over the syllable/mora in question. Such cases were not retrieved in this query. However, by checking for a high value of the gradience parameter for the relevant transition, it is in principle possible to examine this issue with these same stylizations.
The alignment (horizontal/temporal) dimension is more complicated. In theory, it would have been possible to apply the Japanese definition as-is, whereby a phrasal H- should occur on the first mora if the initial syllable is heavy and the second mora if the initial syllable is light (cf. §4.1.1). This would have required (1) distinguishing between initial heavy vs. light syllables, and (2) if the initial syllable is light, determining where the second mora is. Both of these issues require making assumptions about what constitutes a heavy vs. light syllable in the interlanguage of these learners. (Should weight be defined based on the L1 or the L2? Would all learners pattern the same in this regard?) The picture is complicated further by variability in loanword adaptation. (For example, for the purposes of weight calculation, should 'pizzeria' be rendered as \textit{pi.za.ri.a} or \textit{pi.t.isa.ri.a}? Likewise, for 'Vietnamese', \textit{be.to.na.mi.zu} or (somewhat less likely) \textit{bi.e.to.na.mi.zu}?)

To circumvent the necessity of making dangerously strong assumptions on such matters, the alignment of phrasal H- was defined as occurring anywhere in the first or second syllable. By counting high targets in either of the first two syllables, all relevant cases are guaranteed to be included. Furthermore, since this region is, by design, before the primary stress in the word, a high target anywhere in that region is non-targetlike (given the exclusions discussed below). As such, it is safe to take an open-ended approach and treat high targets on either of the first two syllables as candidates for phrasal H-.

The second stage required clustering the target words into three groups according to their stress patterns, as laid out in the table below: In the 'Stress pattern' column, '1' indicates primary stress, '2' indicates secondary stress, and '0' indicates unstressed. The implementation of the
queries needed to be sensitive to these groups because, for example, the secondary stress falls on the 1st syllable in Group I but the 2nd syllable in Group II.

<table>
<thead>
<tr>
<th>Group</th>
<th>Stress pattern</th>
<th>Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group I: 201(0)(0)</td>
<td></td>
<td>masquerade</td>
</tr>
<tr>
<td></td>
<td>201</td>
<td>catastrophic, ambidextrous, academic,</td>
</tr>
<tr>
<td></td>
<td>2010</td>
<td>continental, artificial, pizzerias, scientific</td>
</tr>
<tr>
<td></td>
<td>20100</td>
<td>agricultural, curiosity</td>
</tr>
<tr>
<td>Group II: 0201(0)(0)</td>
<td></td>
<td>Vietnamese</td>
</tr>
<tr>
<td></td>
<td>0201</td>
<td>cooperation</td>
</tr>
<tr>
<td></td>
<td>02010</td>
<td>encyclopedias</td>
</tr>
<tr>
<td>Group III: 021</td>
<td>021</td>
<td>employee</td>
</tr>
</tbody>
</table>

Table 6.2: Initial words in Research Question 1 sentence classified by stress pattern

The second stage involved using this information to filter out any 'lookalike' categories within the L2 English system that superficially resemble a phrasal H-. This stage of the analysis was implemented in an effort to be conservative and not falsely classify something perfectly targetlike as transfer. These lookalikes took two forms. First, the peak that could be interpreted as a phrasal H- must not have been analyzable as a H* (or L+H*) pitch accent associated with the secondary stress. This takes into account cases where both the primary and secondary stresses receive pitch accents. For many words, this also takes into account stress-shifts triggered by analogy to a Japanese loanword or a high-frequency morphologically underived word (e.g. producing *pizzerias* out of analogy with English 'pizza' or its Japanese loanword form 'piza'.) The following table contains more specific details about how this was operationalized.
<table>
<thead>
<tr>
<th>Group</th>
<th>Secondary stress</th>
<th>Primary stress</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group I: 201(0)(0)</td>
<td>1st σ ends H</td>
<td>2nd σ ends L</td>
</tr>
<tr>
<td></td>
<td>3rd σ begins H</td>
<td></td>
</tr>
<tr>
<td>Group II: 0201(0)(0)</td>
<td>2nd σ ends H</td>
<td>3rd σ ends L</td>
</tr>
<tr>
<td></td>
<td>4th σ begins H</td>
<td></td>
</tr>
<tr>
<td>Group III: 021</td>
<td>1st σ ends L</td>
<td>2nd σ begins H</td>
</tr>
</tbody>
</table>

Table 6.3: Definition of secondary stress (excluded for Research Question 1)

This table can be read as follows. '1st/2nd/3rd σ' refers to a specific syllable within the word. 'begins H/L" means that the first vertex within the specified syllable must be H/L. Conversely, 'ends H/L' means that the last vertex within the specified syllable must be H/L. The definitions for groups I and II describe the same kind of case, whereby (1) the secondary stress has a H* (or a L+H*, since any preceding vertices are unspecified), and (2) the primary stress has a L+H* (as indicated by the '_ ends L, _ begins H' sequence). In other words, these definitions capture cases where the primary and secondary stressed syllables both have pitch accents. Group III is slightly different. Since the secondary and primary stresses are adjacent, the definition is merely that the secondary stress needs to have a L+H*.

Secondly, the peak that could be interpreted as a phrasal H- must not have been analyzable as a H+!H* pitch accent associated with the primary stress. For Group III (the 021 word *employee*), this was simply [2nd σ ends H, 3rd σ begins L] (i.e., the primary stress has a low target immediately preceded by a high). For Groups I and II, a token had to meet both condition (1) and condition (2) below:

(1) a. [X-1 σ ends H] OR ( [X-2 σ ends H] AND [X-1 σ has no vertices] )
    b. The vertex immediately preceding the primary stress is a H and is the last vertex inside one of the two preceding syllables.

(2) a. [X σ begins L] OR ( [X σ has no vertices] AND [X+1 σ begins L] )
    b. An L occurs as the first vertex inside either the primary stress or, where possible, the immediately following syllable.
Versions (1a) and (2a) contain the same formalism used in Table 6.3, with X indicating the primary-stressed syllable. A prose translation of the two conditions is provided in (1b) and (2b). At a more conceptual level, these two conditions attempt to capture the generalization that a low target for the second tone (+!H*) in H+!H* must occur either inside the primary stress or shortly thereafter, and the high target for the first tone (H+) must occur 1-2 syllables before the primary stress.

With the two classes of exclusions thus defined, the query could be run in order to ascertain the frequency of occurrence of the relevant kind of crosslinguistic transfer. The following section describes the results thus obtained.

6.1.2 Results A: Relative frequency overall

The following table contains the percentage of tokens returned by the query described in the previous section, broken apart by individual sentence and the learner-vs.-native distinction. Since each token for a given sentence is produced by a different speaker, this effectively represents what proportion of the natives and learners read each sentence with a contour suggestive of a phrasal H-.
<table>
<thead>
<tr>
<th>#</th>
<th>Stimulus sentence</th>
<th>Learners</th>
<th>Natives</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Cà.ta.stró.phic economic cutbacks neglect the poor.</td>
<td>3/25 (12%)</td>
<td>0/11 (0%)</td>
</tr>
<tr>
<td>2</td>
<td>Am.bi.déx.trous pickpockets accomplish more.</td>
<td>3/25 (12%)</td>
<td>0/11 (0%)</td>
</tr>
<tr>
<td>3</td>
<td>À.ca.dé.mic aptitude guarantees your diploma.</td>
<td>4/25 (16%)</td>
<td>1/11 (9.1%)</td>
</tr>
<tr>
<td>4</td>
<td>Còn.ti.nén.tal drift is a geological theory.</td>
<td>1/25 (4%)</td>
<td>0/11 (0%)</td>
</tr>
<tr>
<td>5</td>
<td>Co.ò.per.á tion along with understanding alleviate dispute.</td>
<td>0/25 (0%)</td>
<td>1/11 (9.1%)</td>
</tr>
<tr>
<td>6</td>
<td>Màs.quer.áde parties tax one's imagination.</td>
<td>10/25 (40%)</td>
<td>1/11 (9.1%)</td>
</tr>
<tr>
<td>7</td>
<td>En.cy.clo.pé.di.as seldom present anecdotal evidence.</td>
<td>0/20 (0%)</td>
<td>1/11 (9.1%)</td>
</tr>
<tr>
<td>8</td>
<td>Àr.ti.fí.cial intelligence is for real.</td>
<td>4/21 (19%)</td>
<td>1/11 (9.1%)</td>
</tr>
<tr>
<td>9</td>
<td>Pl.zze.rí.as are convenient for a quick lunch.</td>
<td>1/22 (4.5%)</td>
<td>0/11 (0%)</td>
</tr>
<tr>
<td>10</td>
<td>Sci.en.tí.fic progress comes from the development of new techniques.</td>
<td>3/21 (14.3%)</td>
<td>1/11 (9.1%)</td>
</tr>
<tr>
<td>11</td>
<td>Àg.ri.cúl.tral products are unevenly distributed.</td>
<td>2/22 (9.1%)</td>
<td>0/11 (0%)</td>
</tr>
<tr>
<td>12</td>
<td>Vi.èt.na.mése cuisine is exquisite.</td>
<td>0/21 (0%)</td>
<td>1/11 (9.1%)</td>
</tr>
<tr>
<td>13</td>
<td>Çú.ri.ó.si.ty and mediocrity seldom coexist.</td>
<td>1/22 (4.5%)</td>
<td>0/11 (0%)</td>
</tr>
<tr>
<td>14</td>
<td>Em.plòy.ée layoffs coincided with the company's reorganization.</td>
<td>0/20 (0%)</td>
<td>0/11 (0%)</td>
</tr>
<tr>
<td></td>
<td><strong>Total:</strong></td>
<td>32/319 (10%)</td>
<td>7/154 (4.6%)</td>
</tr>
</tbody>
</table>

*Table 6.4: Number of tokens exhibiting phrasal H- pattern (in raw counts and percentages)*

The following plot graphically represents the percentage data from the above table. The sentences (i.e. pairs of bars) are sorted based on the size of the difference between the learners and natives (ordered from learners higher to natives higher). The sentence numbers along the bottom of the plot correspond to the numbers in the leftmost column of the table.
As indicated in the rightmost column of the table, there was never more than 1 token returned for the native speakers. (A simple $N=1$ creates such a large spike in the barplot merely because there were only 11 native speaker tokens for each sentence.) Since these native speakers lived in Japan, it is possible these tokens represent influence of the Japanese language on their L1 English production. However, it is somewhat more likely that this represents nothing more than unsystematic noise due to slight imperfections in the design of the query.

The rate for the learners is also relatively modest, with the median percentage of 9.1% across all 14 sentences. Such a low figure suggests that most learners are behaving in targetlike ways over the targeted portion of the F0 contours in their L2 English production. In other words,
transfer of this particular aspect of these learners' L1 intonational phonology appears to be only sporadic in nature.

As can be seen in the barplot, the rate of phrasal H- is markedly highest for sentence #6 (Masquerade parties tax one's imagination.). (For a visualization what the data for this sentence looks like, refer back to Figure 6.2 above.) It is possible that the especially high rate for this particular sentence is tied to the fact that this is the only word under examination with a '201' stress pattern (cf. Table 6.2). However, upon manually inspecting several of the tokens returned by the query for this sentence, alternative analyses (not involving phrasal H-) were possible for several of them. The same is also true for a handful of other tokens returned by the query for other sentences. Thus, if anything, the actual median rate of phrasal H- occurrence may be somewhat lower than the 9.1% reported above.

While relatively few in number, many of the tokens returned by the query do indeed appear to contain instances of phrasal H-. The following is one such example.
Figure 6.4: Example of a learner token of sentence 2 (Ambidextrous...) with a phrasal H-

In this example, the F0 first rises to a plateau during the [m] before rising yet again to a peak on the [ɛ]. Using the set of possible contours afforded by the stylization approach in Chapter 3, it is impossible to satisfactorily capture this with a single F0 transition. It indeed appears this learner is rising to a distinct F0 target during the first syllable of the utterance - exactly where a phrasal H- is predicted to appear. A similar observation can also be made for the following example.
Figure 6.5: Example of a learner token of sentence 8 (Artificial...) with a phrasal $H$-

The story here is similar to the one just told. Here, the purported phrasal $H$- occurs during the [ә], followed by a rise to an even higher value on [ɪ] (in the primary-stressed syllable). While the rise preceding the peak on a primary stress is often described as L+H*, the relevant target in this case is clearly not (paradigmatically) low, being well above the baseline for this speaker's F0 range (around 180 Hz). Thus, it appears the two F0 targets in question are Japanese $H$- followed by English $H^*$ - an interlanguage mixing of intonation categories from two languages inside a single intonation contour.

6.1.3 Results B: Frequency predicted by ratings

The analyses presented above did not take into account learners' proficiency in L2 phonology. In other words, all learners at all levels were pooled together in the analysis. To fill
in this gap, the present section explores the extent to which the rating data (a proxy for proficiency) can predict the appearance of a phrasal H- in the dataset described above.

With the learners divided into three groups (high-rated, low-rated, and mid) for each rating criterion as spelled out in Chapter 5 (§5.3.5), the rate of phrasal H- appearance could be calculated separately for each of the groups. For example, note that there are a total of 319 data points for the learners in Table 6.4. After excluding 12 tokens produced by the 7 learners who were not rated, the remaining 307 data points both have (1) a three-way classification into High-Mid-Low as well as (2) a binary 'Did this token display the relevant phenomenon (for Research Question 1, a phrasal H-)'? By integrating these two pieces of information, separate rates of phrasal H- appearance could be calculated for each combination of (pseudo-)proficiency level and rating criterion. 3

The results of applying this analysis to the Research Question 1 dataset are presented out below.

---

3 Note that piece of information (2) is determined on a token-by-token basis. As such, while the imbalance among the 8 stimulus lists in the selected sentences (as documented in §5.3.5) means that certain learners contribute to the overall group-level rate more than others, token-to-token variation within individual learners has the potential to wash out any concomitant skewing in the data.
Binomial tests of the rate of phrasal H- appearance between high-vs-low groupings are not significant for any of the five rating tasks. The flatness of these results are presumably due in no small part to the fact phrasal H- is relatively rare in the dataset as a whole.

6.1.4 Revisiting Research Question 1

As posted in Chapter 4 (§4.2.1), Research Question 1 is as follows:

(3) **Research Question 1**: In their L2 English production, how often do Japanese EFL learners exhibit evidence for an L1-transferred phrasal H-?

The overall answer to this question appears to be *at a median rate of 9.1% of the utterance tokens for any given sentence*. Moreover, learners who are rated more highly do not have any significantly lower rates of phrasal H- appearance. Thus, it seems phrasal H- appearance may be best conceived of as a constant residual phenomenon, cutting across proficiency levels.
6.2 Research Question 2

Research Question 2 seeks evidence for disproportionately frequent use of low phrase-final edge tones in Japanese EFL learners' production, suggesting transfer of Japanese L%. It is important to keep in mind that, unlike Research Question 1, the transfer in question here does not stem from a black-and-white case of L1 (Japanese) having some feature whereas the L2 (English) does not. Rather, the difference is one of overall proportionality, such that Japanese has the feature *more frequently* than English. As such, the query is expected to return tokens for both learners and native speakers, but the returned tokens should be much more numerous for the learners.

Before proceeding to describe the query itself, two methodological issues require comment. Since this query pertains to edge tones, it requires first establishing a clear definition to operationalize the construct of what counts as a 'prosodic boundary'. The approach adopted here is to base the definition on certain specific classes of silences that were explicitly marked in the segmentation stage of data processing. By definition, this approach fails to catch more minor prosodic boundaries with no accompanying silence at all. (Note that ToBI systems, e.g. X-JToBI (Maekawa et al. 2002), generally have separate break index labels depending on whether there is a pause.) However, recall that silences of any length were marked in the segmentation, even those of short duration. This fact is displayed in the following plot, which shows the distribution in silence duration across all tokens analyzed for Research Question 2. The overall median of the distribution is 89 ms.
Figure 6.7: Distribution in silence duration (x-axis) on a (natural) log scale

Since even very short silences were marked, tying the definition of a prosodic boundary to silent portion of the signal should catch even relatively minor prosodic boundaries. Only a small fraction of boundaries (with no silence whatsoever and only tonal marking) will not be yielded by the query.

A second, related methodological issue concerns the fact that utterances produced by learners inherently have systematically more silent portions than the natives, irrespective of what the speakers did with intonation. The following figure attests to this fact.
Each of the two panels represents a histogram of the total number of silences marked inside the segmentation of each individual token. (The y axis is not standardized between the two plots in order to reflect the fact that, as a learner corpus, the ERJ contains more tokens overall for the learners.) The fact that learners have more silences is most likely tied to limitations in fluency (e.g. slower reading rates) and/or on-line processing (e.g. slower syntactic parsing), both of which reflect universal/developmental factors rather than crosslinguistic transfer per se. While the ERJ (including the set of learners under examination) contains many advanced learners, many others were of low enough proficiency that they struggled to read the text aloud smoothly. For such learners, rather than the intended sentence reading task per se, the task may have been more just a matter of identifying and pronouncing individual words one at a time.

Critically, however, not all silences are marked intonationally. Many of the silences marked in the segmentation were merely short silences preceding obstruents (as evidenced by the low tail of the distribution in Figure 6.7). Even a longer silence (i.e. a true 'pause') is not
guaranteed to have intonational marking, such as in the case of minor disfluencies where an F0 transition simply 'picks up where it left off' before the silence. Furthermore, *not all intonation-marked boundaries are preceded by a low F0 target*. Many utterance-medial prosodic boundaries are marked by a boundary rise - i.e. a high F0 target rather than a low one. Other prosodic boundaries are signaled through F0 range reset alone, in which case, again, there would not be any low F0 target before the silence. Thus, the mere presence of a silence does not guarantee the phenomenon under examination in Research Question 2 will occur. As such, even though low-proficiency learners may have an increased rate of pausing (due to fluency and processing constraints), the inclusion of such learners in the dataset does not compromise the validity of the analysis.

With these important methodological issues thus clarified, the following section details the specific architecture for the Research Question 2 query.

**6.2.1 Structure of query**

As with Research Question 1, the time domain of the stylizations was first standardized to percentages across the segmentation intervals. The general details of this time-normalization were the same as that described above for Figure 6.2. Here again, this information was used to determine the 'affiliation' of each vertex (i.e. which syllable the vertex should be counted as 'belonging to'). While syllable affiliation cutoffs were determined on a sentence-by-sentence basis for Research Question 1, no similar process was applied here. Rather, each vertex was straightforwardly assigned to the syllable in which it fell.\(^4\)

---

\(^4\) In the occasional case where a vertex fell inside an interval marked as a silence, it was treated as associated with the nearest non-silent interval. For example, if a vertex occurred at 90% across a silent interval, it was treated as...
With each vertex thus assigned to a specific syllable in the utterance, the vertices could be coded according to a classification scheme, as described in the following table:

<table>
<thead>
<tr>
<th>#</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Q1 Is this the last vertex inside a given interval?</td>
</tr>
<tr>
<td></td>
<td>Q2 Is this vertex inside an interval for a silence?</td>
</tr>
<tr>
<td></td>
<td>Q3 Is this vertex inside an interval that is immediately followed by a silence?</td>
</tr>
<tr>
<td>B</td>
<td>Q4 If Q2 or Q3 is TRUE, is that silence word-medial?</td>
</tr>
<tr>
<td></td>
<td>Q5 If Q2 or Q3 is TRUE, does that silence occur before a stop or affricate?</td>
</tr>
<tr>
<td></td>
<td>Q6 If Q2 or Q3 is TRUE, is that silence over 200 milliseconds?</td>
</tr>
<tr>
<td>C</td>
<td>Q7 Is the transition to this vertex a syntagmatic rise from the preceding vertex?</td>
</tr>
<tr>
<td></td>
<td>Q8 Is the transition to this vertex under 50 cents in magnitude?</td>
</tr>
<tr>
<td>D</td>
<td>Q9 Is the current vertex inside the interval for a monosyllabic function word?</td>
</tr>
<tr>
<td></td>
<td>Q10 What is the metrical level for the interval that contains the current vertex?</td>
</tr>
</tbody>
</table>

Table 6.5: Ten questions asked of every vertex

Collectively, these ten questions provide all of the information upon which the query for Research Question 2 is based. Via looping in a script, all ten questions are 'asked' separately for every individual vertex in every stylization. In most cases, the 'answer' is boolean/logical (either TRUE, FALSE, or N/A if not applicable). The only exception is Q10, for which the answer is an integer (1 or 2 or 0 for primary/secondary/unstressed).

At a higher conceptual level, the questions are broken into four groups (indicated with 'A' through 'D' in the above table). First, the query must locate vertices that have some associated silence (Group A). Second, that silence must meet certain criteria for position, duration, and consonantal context (Group B). Third, the vertex in question must represent a low F0 target affiliated with whatever syllable followed that silence (and occurring at 0% across its duration). The same process was also applied in the few rare cases where vertices were located before the first segmentation interval, or after the last segmentation interval, of the utterance as a whole.
(Group C). Fourth, the vertex should not be analyzable as a pitch accent such as L* (Group D).

Further details about each of the questions is provided below, broken into these four groups.

Question Group A asks, for every vertex in every stylization, whether it has some 'associated' silence. Here, 'associated silence' is defined disjunctively as either one of two scenarios. First, the vertex could be TRUE for Q2, indicating that it occurs inside a silent interval. Such cases occurred relatively infrequently since silent intervals generally do not contain vertices. The second, more typical scenario is where the vertex is TRUE for both Q1 and Q3, indicating that the vertex occurs inside a normal (non-silent) syllable, but it is the last vertex inside this syllable and this is followed by a silence. This describes a more prototypical kind of L%, where the low target occurs inside the last syllable before a silence. Since the segmentation never contained two adjacent silences, any given vertex can only return TRUE for Q2 or Q3 (not both). Also note that the portion of the soundfile after the end of the utterance was not marked as a silence and is therefore not picked out by these questions. As such, utterance-final boundaries are excluded from the scope of the analysis.

Question Group B checks whether any associated silence determined in Group A meets certain criteria for position, length, and consonantal context in order to qualify as a true 'pause'. First, Q4 identifies whether the silence associated with a vertex occurs at a syllable boundary word-medially (e.g. engi-...-neering) as opposed to between two words (e.g. engineering ... department). Word-medial silences were presumed to be disfluencies rather than bona fide pauses. Second, Q5 checks whether the silence occurred before a non-continuant consonant - more specifically, a voiceless stop ([p], [t], or [k]), a voiced stop ([b], [d], or [g]), or an affricate ([dʒ]). Since such consonants inherently create periods of silence, Q5 performs the crucial
function of singling these cases out to be treated specially. The exact syllables picked out by this question are listed in the following table. Numbers in parentheses correspond to the sentence numbers provided in the Research Question 2 segmentation table in §5.4.2.

<table>
<thead>
<tr>
<th>Segment</th>
<th>Syllable (sentence number)</th>
</tr>
</thead>
<tbody>
<tr>
<td>[p]</td>
<td>part (1&amp;7), -pen (2), -pub- (3), Pa- (3), pine (4), -ppoint- (6), -part- (6), picked (8), pairs (8), plan (10), pre- (11), -pared (11), -pper (11), -py (12)</td>
</tr>
<tr>
<td>[t]</td>
<td>to (2&amp;3), trees (4), -time- (7), -ting (7), -tem (7), time (9), -times (12), team (13)</td>
</tr>
<tr>
<td>[k]</td>
<td>caught (1), cost (2), co- (3&amp;7), -quoi- (4), care- (5), -come (7), -cause (10), caused (12), cured (12), coach (13), -cons (14)</td>
</tr>
<tr>
<td>[b]</td>
<td>by (2&amp;12), bear (3), be (5&amp;9), bi- (6), break- (7), bro- (8), best (9), busi- (9), be- (10), -ball (13), black (14)</td>
</tr>
<tr>
<td>[d]</td>
<td>de- (6), -ded (7), -down (7), di- (11), dime (13)</td>
</tr>
<tr>
<td>[g]</td>
<td>Gus (4), group (12)</td>
</tr>
<tr>
<td>[dʒ]</td>
<td>job (2), -jec- (3), joint (6), -gy (6), -gi- (6&amp;10), -gers (9), -dget (10)</td>
</tr>
</tbody>
</table>

Table 6.6: List of syllables beginning with non-continuant consonants (stops and affricates), spelled based on source word, cross-referenced with sentence numbers from §5.4.2

Finally, Q6 factors length into the equation by identifying which vertices have an associated pause that is over a predetermined length threshold. In the present context, this cutoff was set at 200 ms - a value determined based on inspecting distributions like that in Figure 6.7. Note that this is over two times the median of the distribution silence length mentioned earlier (89 ms).

For all vertices with an associated silence (as determined by question group A), Q4, Q5, and Q6 effectively carve out a 2x2x2 typology of eight different sub-types of silence. The following row sketches out this typology, grouped into four clusters ((a) through (d)). The N for each sub-type is provided in the bottom row.
Table 6.7: Eight different sub-types of silence, as defined by Q4, Q5, and Q6

Of the four groups, type (a) was excluded from the definition of true 'pause' because both involve short silences before a non-continuant (stop or affricate) consonant, which are presumably merely a byproduct of the consonant. Type (b), representing anything else that is word-medial, was excluded for presumably being disfluencies. (The crucial cells for both (a) and (b) are indicated in bold.) The other two types were retained: (c) long silences before a non-continuant, and (d) silences in other environments of any length. Thus, Q4, Q5, and Q6 function to define any kind of silence falling into either of these categories as constituting a 'pause'.

Question Group C ascertains whether the vertex in question represents a low F0 target. This portion of the query is relatively straightforward. Q7 defines the rise vs. fall distinction, and Q8 filters out transitions that are too flat (using the 50-cent criterion discussed above for Research Question 1). Any vertex occurring at the endpoint of a falling F0 transition of large enough magnitude was classified as low.

In the effort to be conservative, question Group D ensures the vertex should not be analyzable as a pitch accent (e.g. L*). To do so, first, in order to implement the observation that function words are normally not pitch-accented in English, Q9 draws a distinction between function words and content words. Here, a function word is defined as any monosyllabic word that is not an adjective/adverb (fresh/strange/thin), proper name (Al/Gus/Ralph), noun
(dime/sauce/socks), or verb (caught/picked/think). In terms of the specific words making up the sentences for Research Question 2, this definition selects the following:

(4) a. **Conjunction**: and, if
   b. **Copula**: be, is
   c. **Determiner**: a, each, the, these, those
   d. **Modal**: may, should, will
   e. **Negation**: not
   f. **Preposition**: as, by, for, in, of, on, through, to, up, with
   g. **Pronoun**: he, him, his, I, them, we, you

Along similar lines, Q10 identifies the stress status (primary, secondary, unstressed) of the syllable where the vertex falls. This information can then be used to implement the observation that pitch accents are normally associated with the primary-stressed syllable in a word. The following information was used to identify the stress status in this way; any syllable indicated with a 1 in the 'Stress' column was treated as primary-stressed. Italicization of a word in the first row indicates it also belongs to the class of function words identified in (4). Bold indicates primary stress.
Table 6.8: All words in the sentences for Research Question 2 divided by number of syllables ("σ") and stress pattern

Q9 and Q10 are then integrated to single out possible cases where the low F0 target is associated with a pitch accent. More specifically, such cases are defined as any syllable that Q9 identifies as not being a monosyllabic function word and Q10 identifies as being a primary stress (’1’). Any low F0 target occurring on such a syllable is excluded from the results of the query. Anything else is included, i.e. both a vertex on a monosyllabic function word as well as one on a secondary-stressed or unstressed syllable in a content word.

To summarize, the query for Research Question 2 consists of 10 'questions' asked of every vertex in every stylization. Group A ensures that the vertex has some associated silence (to constrain the results to F0 targets occurring before prosodic boundaries). Group B factors out short, word-medial, and/or pre-obstruent silences in order to leave only 'true' pauses. Group C determines whether the F0 target is syntagmatically low. Finally, Group D brings in external information about what constitutes a content vs. function word and metrical stress levels in order

<table>
<thead>
<tr>
<th>σ</th>
<th>Stress</th>
<th>Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>a, Al, and, as, be, bear, best, black, by, caught, caused, coach, cost, cured, dime, each, eight, first, for, fresh, group, Gus, has, he, him, his, I, if, in, is, job, joint, lack, long, loss, may, nine, not, now, of, on, pairs, part, picked, pine, plan, Ralph, red, rich, right, sauce, saw, shell, shock, shoes, should, socks, state, store, strange, team, the, them, these, thin, think, those, through, time, to, trees, up, walk, watch, we, will, with, you</td>
</tr>
<tr>
<td>2</td>
<td>01</td>
<td>because, instead, invest, prepared</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>answers, brother, budget, business, dinner, failure, forest, heating, lemon, mergers, open, Paris, received, revised, shrapnel, snapper, stylish, system, zircons</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>breakdown, football, income, inmost, insect, redwoods, sometimes</td>
</tr>
<tr>
<td>3</td>
<td>010</td>
<td>appointment, departments, republic, sequoia, subjection</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>carefully, national, therapy</td>
</tr>
<tr>
<td></td>
<td>102</td>
<td>commonwealths</td>
</tr>
<tr>
<td></td>
<td>210</td>
<td>straightforward, untimely</td>
</tr>
<tr>
<td>4</td>
<td>0100</td>
<td>biology, original</td>
</tr>
<tr>
<td></td>
<td>2010</td>
<td>coincided, engineering</td>
</tr>
</tbody>
</table>
to filter out low F0 targets associated with pitch accents. The following section describes the
results obtained when the query thus designed was run on the stylization and segmentation data
for Research Question 2.

6.2.2 Results A: Relative frequency overall

The following table describes the number of instances of the phenomenon in question
(low boundary marking, i.e. L1 Japanese L% or L2 English L-) on a sentence-by-sentence basis.
The 'Learners' and 'Natives' cell in the row for each sentence represents the overall rate of
occurrence of low boundary marking for all speakers combined. For example, in the case of the
[Sentence 5, Learners] cell, "1.19 (32x in 27)" means that across the 27 learner tokens for that
sentence, low boundary marking occurred a total of 32 times, hence a rate of 1.19 instances per
sentence.
<table>
<thead>
<tr>
<th>#</th>
<th>Stimulus sentence</th>
<th>Learners</th>
<th>Natives</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>I caught a strange insect in the inmost part of the forest.</td>
<td>0.86 (18x in 21)</td>
<td>0.55 (6x in 11)</td>
</tr>
<tr>
<td>2</td>
<td>His failure to open the store by eight cost him his job.</td>
<td>0.26 (6x in 23)</td>
<td>0.36 (4x in 11)</td>
</tr>
<tr>
<td>3</td>
<td>These commonwealths will not long bear a state of subjection to the republic of Paris.</td>
<td>0.77 (20x in 26)</td>
<td>0.10 (1x in 10)</td>
</tr>
<tr>
<td>4</td>
<td>Gus saw pine trees and redwoods on his walk through Sequoia National Forest.</td>
<td>0.81 (21x in 26)</td>
<td>0.09 (1x in 11)</td>
</tr>
<tr>
<td>5</td>
<td>Those answers will be straightforward if you think them through carefully first.</td>
<td>1.19 (32x in 27)</td>
<td>0.40 (4x in 10)</td>
</tr>
<tr>
<td>6</td>
<td>Al received a joint appointment in the biology and the engineering departments.</td>
<td>1.04 (28x in 27)</td>
<td>0.40 (4x in 10)</td>
</tr>
<tr>
<td>7</td>
<td>Each untimely income loss coincided with the breakdown of a heating system part.</td>
<td>1.12 (27x in 24)</td>
<td>0.09 (1x in 11)</td>
</tr>
<tr>
<td>8</td>
<td>He picked up nine pairs of socks for each brother.</td>
<td>0.68 (17x in 25)</td>
<td>0.36 (4x in 11)</td>
</tr>
<tr>
<td>9</td>
<td>Right now may not be the best time for business mergers.</td>
<td>0.12 (3x in 25)</td>
<td>0.00 (0x in 11)</td>
</tr>
<tr>
<td>10</td>
<td>We revised the original plan because of the lack of the budget.</td>
<td>0.67 (16x in 24)</td>
<td>0.00 (0x in 11)</td>
</tr>
<tr>
<td>11</td>
<td>Ralph prepared red snapper with fresh lemon sauce for dinner.</td>
<td>0.14 (4x in 28)</td>
<td>0.09 (1x in 11)</td>
</tr>
<tr>
<td>12</td>
<td>Shell shock caused by shrapnel is sometimes cured through group therapy.</td>
<td>0.78 (21x in 27)</td>
<td>0.30 (3x in 10)</td>
</tr>
<tr>
<td>13</td>
<td>The football team coach has a watch as thin as a dime.</td>
<td>0.14 (4x in 28)</td>
<td>0.09 (1x in 11)</td>
</tr>
<tr>
<td>14</td>
<td>The rich should invest in black zircons instead of stylish shoes.</td>
<td>0.89 (25x in 28)</td>
<td>0.09 (1x in 11)</td>
</tr>
</tbody>
</table>

All sentences pooled: 0.67 (242x in 359) 0.21 (31x in 150)

Table 6.9: Rate of low boundary marking per sentence ([Number of times] in [number of sentences])

The following plot graphically represents the rate data from the above table. As was done above for Research Question 1, the sentences (i.e. pairs of bars) are sorted based on the size of the difference between the learners and natives (ordered from learners higher to natives higher). The sentence numbers along the bottom of the plot correspond to the numbers in the leftmost column of the table.
Figure 6.9: Rate of low boundary marking per sentence

This plot makes it clear that, as a whole, learners systematically show more low boundary marking than native speakers, as predicted. This pattern shows up clearly in all but the rightmost four sentences (#2, #9, #11, and #13), where the rates of low boundary marking were low overall. This is perhaps because these four sentences were among the shortest of all the sentences (only 13-14 syllables long), alleviating the need to insert prosodic boundaries. (Only sentence 8 is shorter, at 11 syllables.)

The above analysis collapsed all tokens of each sentence together. The following plot shows the distribution of how many instances of low boundary marking occurred inside any one individual token. (Only tokens with 1 or more instance are represented in the plot.)
Figure 6.10: Rate of low boundary marking per token

With the exception of 1 token with two instances, all other 29 native tokens with low boundary marking have it only a single time. The learner distribution is much more diffuse. In addition to 117 tokens with one instance and 48 tokens with two, there are also 8 tokens with three and even 1 token with five. This means that, whereas a native speaker would generally have only one boundary, e.g. \{ *I caught a strange insect* \} \{ *in the inmost part of the forest* \}, a learner would break an equivalent sentence into two, three, or even more smaller pieces. These same data can also be described in terms of the percentage of N=1 boundary per token.

When looked at in this way, the contrast is striking: 96.7% (29/30) for native speakers versus 67.2% (117/174) for learners. This means that nearly one third (32.8%) of all learner tokens with low boundary marking exhibit it more than once. Thus, in the utterances where Japanese EFL learners do low boundary marking, they frequently do it more than once. This reinforces the evidence for the pervasiveness of this phenomenon in this learner population.
The third and final analysis dichotomizes the number of instances per token as 0 vs. 1 or more. (Thus, in this analysis, a token with 1 instance of low boundary marking would be treated the same as a token with 5.)

<table>
<thead>
<tr>
<th>#</th>
<th>Stimulus sentence</th>
<th>Learners</th>
<th>Natives</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>I caught a strange insect in the inmost part of the forest.</td>
<td>14/21 (66.7%)</td>
<td>6/11 (54.5%)</td>
</tr>
<tr>
<td>2</td>
<td>His failure to open the store by eight cost him his job.</td>
<td>5/23 (21.7%)</td>
<td>4/11 (36.4%)</td>
</tr>
<tr>
<td>3</td>
<td>These commonwealths will not long bear a state of subjection to the republic of Paris.</td>
<td>14/26 (53.8%)</td>
<td>1/10 (10%)</td>
</tr>
<tr>
<td>4</td>
<td>Gus saw pine trees and redwoods on his walk through Sequoia National Forest.</td>
<td>18/26 (69.2%)</td>
<td>1/11 (9.1%)</td>
</tr>
<tr>
<td>5</td>
<td>Those answers will be straightforward if you think them through carefully first.</td>
<td>20/27 (74.1%)</td>
<td>4/10 (40%)</td>
</tr>
<tr>
<td>6</td>
<td>Al received a joint appointment in the biology and the engineering departments.</td>
<td>18/27 (66.7%)</td>
<td>4/10 (40%)</td>
</tr>
<tr>
<td>7</td>
<td>Each untimely income loss coincided with the breakdown of a heating system part.</td>
<td>17/24 (70.8%)</td>
<td>1/11 (9.1%)</td>
</tr>
<tr>
<td>8</td>
<td>He picked up nine pairs of socks for each brother.</td>
<td>13/25 (52%)</td>
<td>3/11 (27.3%)</td>
</tr>
<tr>
<td>9</td>
<td>Right now may not be the best time for business mergers.</td>
<td>3/25 (12%)</td>
<td>0/11 (0%)</td>
</tr>
<tr>
<td>10</td>
<td>We revised the original plan because of the lack of the budget.</td>
<td>12/24 (50%)</td>
<td>0/11 (0%)</td>
</tr>
<tr>
<td>11</td>
<td>Ralph prepared red snapper with fresh lemon sauce for dinner.</td>
<td>4/28 (14.3%)</td>
<td>1/11 (9.1%)</td>
</tr>
<tr>
<td>12</td>
<td>Shell shock caused by shrapnel is sometimes cured through group therapy.</td>
<td>16/27 (59.3%)</td>
<td>3/10 (30%)</td>
</tr>
<tr>
<td>13</td>
<td>The football team coach has a watch as thin as a dime.</td>
<td>2/28 (7.1%)</td>
<td>1/11 (9.1%)</td>
</tr>
<tr>
<td>14</td>
<td>The rich should invest in black zircons instead of stylish shoes.</td>
<td>18/28 (64.3%)</td>
<td>1/11 (9.1%)</td>
</tr>
</tbody>
</table>

Total: 174/359 (48%) 30/150 (20%)

Table 6.10: Percentage of tokens for each sentence with 1 or more instance of low boundary marking

211
Here again, the same four short sentences (#2, #9, #11, and #13) are the only exceptions to the general trend that learners have a much higher rate of low boundary marking. Expressed in terms of the medians calculated across the 14 sentences, 56.55% of learners used low boundary marking, as opposed to just 9.54% of native speakers.

The following examples illustrate the difference between learners and native speakers. Both examples are of sentence #6 ("Al received a joint appointment in the biology and the engineering departments."). Due to the length of the files, they are each broken into two separate rich visualization plots. The first utterance is from a Japanese EFL learner. This is the token referred to in Figure 6.10 above as having as many as 5 instances of low boundary marking in the same utterance. As can be seen in the segmentation (bottom pane of the plot), there are many silences in this utterance, but only a subset of them are returned by the query.
Figure 6.12: Example learner token with many low F0 targets at phrase boundaries (Part 1)
The following is a native speaker's production of the same sentence. Despite having roughly the same number of silences (11 for the learner vs. 8 for the native), this time there is only a single instance of low boundary marking (at the end of the first half of the utterance).
Figure 6.14: Example native token with few low F0 targets at phrase boundaries (Part 1)
6.2.3 Results B: Frequency predicted by ratings

The results of integrating the rating data into the analysis in a parallel way to that done in §6.1.3 are presented in the following figure:
Figure 6.16: Rate of low boundary marking (y-axis) as predicted by rating group

Comparison with the equivalent plot in §6.1.3 confirms that the phenomenon of low boundary marking is indeed strikingly more frequent than phrasal H- appearance. (Note that the y axes are standardized between the two plots.) Like the phrasal H- results, though, the rating data are generally a poor predictor of the occurrence of low boundary marking. In pairwise comparisons of the Low vs. High groups in the five panels of the above figure, the only significant difference to emerge is for Sentence Rhythm, as revealed by a standard binomial test contrasting 53/116 vs. 59/103 (thus, 45.7% vs. 57.3%). However, even here, the difference is in the opposite direction than expected: the higher-rated learners have more low boundary marking. To the extent this is a robust and meaningful finding, it may mean that an increase in the use of low edge tones, as encouraged by the distribution of L% in the L1, may contribute to the perceived 'rhythmicness' of speech.

This one exception notwithstanding, the overall lack of explanatory power for the rating data suggests that, in this population of Japanese EFL learners, an increased propensity to use
low edge tones (relative to a native speaker baseline) is something shared in common between beginning and advanced learners. While the rating data only indirectly index overall proficiency in L2 phonology, the fact that High learners had equal (if not higher) rates of low boundary marking compared to Low learners for all five of the rating criteria is compelling evidence that this phenomenon is pervasive even in more advanced learners. Recall from the discussion of methodological issues at the very beginning of §6.2 that the fact that, due to processing constraints, low-proficiency learners are generally more disfluent (and therefore produce more silences) is a complication for the present study. The fact that even advanced learners use low boundary marking just as frequently suggests that the use of low boundary marking is not merely a byproduct of the propensity to pause.

6.2.4 Revisiting Research Question 2

As formulated in Chapter 4 (§4.2.2), Research Question 2 is as follows:

(5) **Research Question 2**: In their L2 English production, how often do Japanese EFL learners use low F0 targets at prosodic boundaries?

The above analyses demonstrated that the rates of occurrence are not only relatively high (median 56.55% in Figure 6.11), but also more occur per sentence (cf. 32.8% of tokens having 2 or more in Figure 6.10). Moreover, the integration of the rating data analysis suggests that this phenomenon affects low-rated and high-rated learners to a roughly equal extent. Thus, the proliferation of low boundary marking appears to be an especially ubiquitous phenomenon in this population of learners, cutting across multiple proficiency levels.

In terms of systemic differences sketched out in §4.2.2, it appears these Japanese EFL learners structured their English utterances into the Japanese equivalent of minor phrases, which
is normally not tonally marked in English. Transferring this phrase structure into English brought along with it (as a secondary byproduct) the epenthesis of many extra L% tones, leading to the inflated rates relative to native speakers.

6.3 Research Question 3

Research Question 3 asks about the alignment of the F0 targets making up utterance-final boundary rises. In the single-rise contour, Japanese EFL learners were predicted to rise over the final syllable rather than the nuclear-stressed syllable) due to the influence of their L1 L% category. As mentioned in §5.4.3, for Research Question 3, the final two F0 transitions of the utterance were stylized - the final rise itself plus the immediately preceding transition. As such, the raw data points flowing into this analysis are the final three vertices in the contour, regardless of where they occurred.

6.3.1 Structure of query

Since a small subset of the utterances for Research Question 3 were produced without a boundary rise, these first needed to be identified and set aside. As such, the first step in the analysis was to single out the contours that had boundary rises in the first place. Toward this end, a histogram of the F0 change for the last two transitions across all tokens is displayed in the following two figures. For the final transition (but not the penultimate transition), the distribution is clearly skewed toward the positive end (i.e. rising F0) due to the nature of the materials (yes-no questions).
As can be seen from the figures, relatively few tokens fall within the range of -50 to +50 cents (i.e. the criterion used in Research Questions 1 and 2). To err on the side of simplicity, and since 0 neatly divides both bimodal distributions, 0 was used as the cutoff between 'rising' and
'not rising'. Thus, an increase in F0 of any magnitude was treated a rise for the purposes of the analysis.

Using this criterion, the following table describes the percentage occurrence of final rises, broken apart by individual sentence and the learner vs. native distinction. Since the majority of tokens included final rises, to make the patterns clearer, the results are presented in terms of how large the residual portion of non-rising tokens are. The rows are sorted by the percentages in the "Learners" column, in decreasing order.

<table>
<thead>
<tr>
<th>#</th>
<th>Sentence text</th>
<th>Learners</th>
<th>Natives</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>Will you please confirm government policy regarding waste removal?</td>
<td>10/24 (41.7%)</td>
<td>1/11 (9.1%)</td>
</tr>
<tr>
<td>15</td>
<td>Do you know this man in this photograph?</td>
<td>11/28 (39.3%)</td>
<td>0/11 (0%)</td>
</tr>
<tr>
<td>6</td>
<td>Would you allow acts of violence?</td>
<td>7/23 (30.4%)</td>
<td>0/11 (0%)</td>
</tr>
<tr>
<td>11</td>
<td>Can the agency overthrow alien forces?</td>
<td>6/23 (26.1%)</td>
<td>0/11 (0%)</td>
</tr>
<tr>
<td>4</td>
<td>Does creole cooking use curry?</td>
<td>4/21 (19%)</td>
<td>0/11 (0%)</td>
</tr>
<tr>
<td>9</td>
<td>May I order a parfait after I eat dinner?</td>
<td>4/26 (15.4%)</td>
<td>0/11 (0%)</td>
</tr>
<tr>
<td>8</td>
<td>Are you looking for employment?</td>
<td>3/22 (13.6%)</td>
<td>1/11 (9.1%)</td>
</tr>
<tr>
<td>7</td>
<td>Do you hear the sleigh bells ringing?</td>
<td>3/23 (13%)</td>
<td>0/11 (0%)</td>
</tr>
<tr>
<td>1</td>
<td>Will Robin wear a yellow lily?</td>
<td>3/26 (11.5%)</td>
<td>0/11 (0%)</td>
</tr>
<tr>
<td>2</td>
<td>Did dad do academic bidding?</td>
<td>2/26 (7.7%)</td>
<td>0/11 (0%)</td>
</tr>
<tr>
<td>5</td>
<td>Did you buy any corduroy overalls?</td>
<td>1/23 (4.3%)</td>
<td>0/11 (0%)</td>
</tr>
<tr>
<td>10</td>
<td>Do you have the yellow ointment ready?</td>
<td>1/26 (3.8%)</td>
<td>0/11 (0%)</td>
</tr>
<tr>
<td>13</td>
<td>Is he a mailman?</td>
<td>0/26 (0%)</td>
<td>1/11 (9.1%)</td>
</tr>
<tr>
<td>3</td>
<td>Did you eat lunch yesterday?</td>
<td>0/25 (0%)</td>
<td>0/11 (0%)</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td><strong>55/342 (16.1%)</strong></td>
<td><strong>3/154 (1.9%)</strong></td>
</tr>
</tbody>
</table>

*Table 6.11: Percentage of tokens for Research Question 3 without a final rise*

It is clear from this table that the native speakers are patterning exactly as expected. Since the selected sentences are yes-no questions, they should all be obligatorily pronounced with a boundary rise, and indeed this was the case (100-1.9=) 98.1% of the time. The learners, on the other hand, patterned slightly differently, with only (100-16.1=) 83.9% of tokens having a final
rise. The individual sentences vary extensively in the extent to which final rises occur, ranging from (a targetlike) 0% in two sentences to as much as 41.7% in one sentence. These results could technically be construed as suggesting that a sizable portion of learners have not yet acquired knowledge of the fact yes-no questions in English require a boundary rise. Given the fact that yes/no questions in Japanese also require a boundary rise, however, it seems more likely that the learners simply failed to recognize the fact that the sentences they were reading were questions in the first place (at least until it was too late). This small residual of non-rising tokens will be excluded from all analyses below. Thus, the overall number of tokens decreased from (154-3=) 151 for the native speakers and (342-55=) 287 for the learners, hence (151+287=) 438 total.

Next, of the remaining tokens (i.e. those with final rises), single rises were distinguished from double rises using the 0 cents criterion cutoff established above. The following table expresses the breakdown into these two categories in terms of the rate of the double-rise contour.
## Table 6.12: Percentages of double rises (i.e. antepenultimate F0 < penultimate F0 < final F0) among tokens with final rises

<table>
<thead>
<tr>
<th>#</th>
<th>Sentence text</th>
<th>Learners</th>
<th>Natives</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Will Robin wear a yellow lily?</td>
<td>7/23 (30.4%)</td>
<td>0/11 (0%)</td>
</tr>
<tr>
<td>2</td>
<td>Did dad do academic bidding?</td>
<td>2/24 (8.3%)</td>
<td>0/11 (0%)</td>
</tr>
<tr>
<td>3</td>
<td>Did you eat lunch yesterday?</td>
<td>8/25 (32%)</td>
<td>8/11 (72.7%)</td>
</tr>
<tr>
<td>4</td>
<td>Does creole cooking use curry?</td>
<td>7/17 (41.2%)</td>
<td>2/11 (18.2%)</td>
</tr>
<tr>
<td>5</td>
<td>Did you buy any corduroy overalls?</td>
<td>5/22 (22.7%)</td>
<td>3/11 (27.3%)</td>
</tr>
<tr>
<td>6</td>
<td>Would you allow acts of violence?</td>
<td>3/16 (18.8%)</td>
<td>1/11 (9.1%)</td>
</tr>
<tr>
<td>7</td>
<td>Do you hear the sleigh bells ringing?</td>
<td>8/20 (40%)</td>
<td>7/11 (63.6%)</td>
</tr>
<tr>
<td>8</td>
<td>Are you looking for employment?</td>
<td>8/20 (40%)</td>
<td>2/10 (20%)</td>
</tr>
<tr>
<td>9</td>
<td>May I order a parfait after I eat dinner?</td>
<td>6/22 (27.3%)</td>
<td>0/11 (0%)</td>
</tr>
<tr>
<td>10</td>
<td>Do you have the yellow ointment ready?</td>
<td>9/25 (36%)</td>
<td>7/11 (63.6%)</td>
</tr>
<tr>
<td>11</td>
<td>Can the agency overthrow alien forces?</td>
<td>4/20 (20%)</td>
<td>0/11 (0%)</td>
</tr>
<tr>
<td>12</td>
<td>Will you please confirm government policy regarding waste removal?</td>
<td>9/14 (64.3%)</td>
<td>7/10 (70%)</td>
</tr>
<tr>
<td>13</td>
<td>Is he a mailman?</td>
<td>17/26 (65.4%)</td>
<td>6/10 (60%)</td>
</tr>
<tr>
<td>14</td>
<td>Do you know this man in this photograph?</td>
<td>5/17 (29.4%)</td>
<td>7/11 (63.6%)</td>
</tr>
<tr>
<td></td>
<td><strong>Total</strong></td>
<td>37/162 (22.8%)</td>
<td>159/308 (51.6%)</td>
</tr>
</tbody>
</table>

This same information is presented graphically in the following figure. (As with the plots for Research Questions 1 and 2 (§6.1.2 and §6.2.2), the sentences are sorted by the magnitude of the difference between learners and natives.)
Figure 6.19: Histogram of F0 change (in cents) from in the final F0 transition across all tokens

This plot makes it clear that double-rise contours are very common for both native speakers and learners, though overall more common for the native speakers (cf. 22.8% vs. 51.6% in the final row of the table). Recall from §4.2.3 that English and Japanese both have ways of generating this kind of contour, either of which is perfectly compatible with asking yes-no questions; hence, it is not surprising the rates are so high. In English, this can represent a L* H-H% contour, with the post-nuclear portion forming a high flat plateau before rising in the final syllable. In Japanese, on the other hand, this contour represents an unaccented phrase with a boundary rise. Depending on the details of alignment, the Japanese EFL learners examined here may be doing one or both of these. Since the focus of Research Question 3 is single rises, the double-rise contours will be excluded from further analysis below. (Since the second rise begins at the utterance-final syllable for both languages in double-rise contours, failure to do so would only artificially inflate the rates of final-syllable alignment in both languages.)
Finally, in order to determine which tokens contained boundary rises confined to the final syllable, the affiliation of the vertices to specific syllables in the utterance needed to be determined. As was done for Research Question 2, the raw values from the segmentation were used rather than introducing any special cutoff values akin to those used in Research Question 1. Since the segmentation was at the level of individual segments, one other point is of note regarding syllable affiliation. Recall from §5.4.2 that, in order to make direct comparisons possible across all learner and native files, segmentation boundaries are marked based on a fully targetlike production of the sentence, and as such, epenthesized vowels are grouped into the interval for the adjacent consonant. This means that if a learner produced *mailman* as 'mail[u]man' (with an epenthesized [u] of some duration), [lu] would be grouped together into a single interval.

To summarize, the 'query' for Research Question 3 identifies all tokens where the final transition is rising (in order to examine boundary rises in the first place) and the penultimate transition is falling (in order to exclude double rises). The syllable affiliation of the vertices can then be queried and used as a predictor variable. The results for the data thus partitioned are presented in the following section.

6.3.2 Results A: Relative frequency overall

The following figure is a visual summary of the data examined for this research question:
[1] Will Robin wear a yellow lily?

[2] Did dad do academic bidding?

[3] Did you eat lunch yesterday?

[4] Does creole cooking use curry?

[5] Did you buy any corduroy overalls?


[7] Do you hear the sleigh bells ringing?

[8] Are you looking for employment?

[9] May I order a parfait after I eat dinner?

[10] Do you have the yellow ointment ready?

[11] Can the agency overthrow alien forces?

[12] Will you please confirm government policy regarding waste removal?

[13] Is he a mailman?

[14] Do you know this man in this photograph?

Figure 6.20: Alignment of penultimate vertex in boundary rises (○=Native, ●=Learner)
This figure represents a composite of 14 scatterplots - one for each of the 14 sentences investigated for this research question. The sentence in question is printed in bold as the main title above each plot. The vertical grey lines indicate segmentation of utterances into individual segments. The single thick line indicates the beginning of the final syllable (the crucial cutoff point in the present context). As above for the 'dot plot' in §6.1, each speaker is represented as a single 'row' in a plot (containing just a single dot here). The dots represent the temporal location of the beginning of the final rise, coded such that unfilled circles represent native speakers and filled circles represents Japanese EFL learners. As before, the horizontal location of each dot is based on the location of the relevant vertex in the stylization for that token, time-normalized based on the segmentation for that same token. If the boundary rise began before the final word, the dot is placed on the leftmost grey line. The vertical location of the dots is arbitrary and reflects only subject ID number.

It is apparent from this visualization that, in the majority of the sentences, the English native speakers (unfilled circles) align the beginning of the boundary rise to the vowel inside the nuclear stressed syllable. Due to the design of these materials, this always occurs before the beginning of the final syllable (the thick grey vertical line). The few outliers that run counter to this trend represent cases where the sentence was interpreted and/or read a different way than expected, e.g. with a H* L- H% contour or with early focus (such as "Would you allow acts of violence?").

The Japanese EFL learners (filled circles) pattern somewhat differently. In most panels of the plot, two distinct clusters can be discerned: one to the left of the grey vertical line and another to the right. The cluster to the left represents the the learners who display alignment
values within the range of native norms. Such learners appear to have acquired targetlike alignment patterns in their L2 English. The second cluster in most panels, to the right of the grey vertical line, represents learners who exhibit the expected transfer pattern. Recall that it was hypothesized that many Japanese EFL learners should begin the final rise from the final syllable due to L1 transfer. This is indeed a pervasive pattern observed in most of the panels. In many cases, this ‘transferring cluster’ is located well after the latest-aligned native speaker.

The following table represents this information numerically.

<table>
<thead>
<tr>
<th>#</th>
<th>Sentence text</th>
<th>Learners</th>
<th>Natives</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Will Robin wear a yellow /lɪ</td>
<td>li/ ?</td>
<td>7/16 (43.8%)</td>
</tr>
<tr>
<td>2</td>
<td>Did dad do academic /bɪ</td>
<td>dɪŋ/ ?</td>
<td>9/22 (40.9%)</td>
</tr>
<tr>
<td>3</td>
<td>Did you eat lunch /jɛs.tɚ</td>
<td>der/ ?</td>
<td>9/17 (52.9%)</td>
</tr>
<tr>
<td>4</td>
<td>Does creole cooking use /kɝ</td>
<td>i/ ?</td>
<td>2/10 (20%)</td>
</tr>
<tr>
<td>5</td>
<td>Did you buy any corduroy /oʊ.vɚ</td>
<td>alz/ ?</td>
<td>10/17 (58.8%)</td>
</tr>
<tr>
<td>6</td>
<td>Would you allow acts of /vɑ.r.oʊ</td>
<td>lɛns/ ?</td>
<td>7/13 (53.8%)</td>
</tr>
<tr>
<td>7</td>
<td>Do you hear the sleigh bells /ɪŋ</td>
<td>mɪj/ ?</td>
<td>3/12 (25%)</td>
</tr>
<tr>
<td>8</td>
<td>Are you looking for /məm.plɔɪ</td>
<td>mənt/ ?</td>
<td>9/11 (81.8%)</td>
</tr>
<tr>
<td>9</td>
<td>May I order a parfait after I eat /dɪ</td>
<td>nɔ/ ?</td>
<td>6/16 (37.5%)</td>
</tr>
<tr>
<td>10</td>
<td>Do you have the yellow ointment /ɛ</td>
<td>di/ ?</td>
<td>1/16 (6.2%)</td>
</tr>
<tr>
<td>11</td>
<td>Can the agency overthrow alien /fɔɪ</td>
<td>sɪz/ ?</td>
<td>4/13 (30.8%)</td>
</tr>
<tr>
<td>12</td>
<td>Will you please confirm government policy regarding waste /ɛ.m.mu</td>
<td>ɪl/ ?</td>
<td>3/5 (60%)</td>
</tr>
<tr>
<td>13</td>
<td>Is he a /mɛɪl</td>
<td>mæn/ ?</td>
<td>8/9 (88.9%)</td>
</tr>
<tr>
<td>14</td>
<td>Do you know this man in this /fʊ.ʊ.tɚ</td>
<td>gɹɛf/ ?</td>
<td>10/12 (83.3%)</td>
</tr>
</tbody>
</table>

Total: 88/189 (46.6%) 12/101 (11.9%)

Table 6.13: Percentage of single-rise tokens where the final rise began within the final syllable

The first column of this table contains the text of the sentence in normal orthography, with the exception of the final word which is represented in broad IPA transcription. The cutoff point used to define the beginning of the final syllable is marked with a '|'. The remaining two columns show what percentage of native speakers and Japanese EFL learners began their final
rises from within that syllable. Put another way, the percentages represent the proportion of cases where the penultimate vertex was aligned after the '|' indicated in the 'Sentence text' column. (The antepenultimate vertex is not taken into account in these calculations.) Note the Ns are somewhat imbalanced due to the fact that the exclusion of the double rises had an uneven effect on the various different individual sentences.

The following barplot visually represents the percentages from the final two columns:

![Figure 6.21: Percentage of single-rise tokens where the final rise began within the final syllable](image)

In these results, the native speakers' percentages are generally low, with over half (N=9) of the sentences being 0%. In contrast, the learners' percentages are generally high - as much as 88.9% in sentence 13 ("Is he a mailman?"). In every case, the learners' percentages are greater than the native speakers. Not surprisingly, the overall percentages reflect this difference, with

---

5 Note that the two sentences discussed in §5.2.3 as being particularly problematic due to loanword accent - sentence 5 (overalls) and sentence 14 (photograph) do not particularly stand out in the results, lying in the middle of the distribution for the learner rates across all the sentences. Thus, it appears that if there is any effect of the mismatch between the English stress location and loanword accent location, it is relatively minimal.
46.6% for learners and 11.9% for native speakers. Thus, there is evidence that a substantial portion of learners are exhibiting the hypothesized pattern, namely that the boundary rise begins within the final syllable, consistent with an L1-transferred L% category.

The following two plots illustrate how the targetlike and the non-targetlike patterns contrast.

![Figure 6.22: Native speaker production of "Can the agency overthrow alien forces?"](image)
Figure 6.23: Learner production of "Can the agency overthrow alien forces?"

The first plot represents a native speaker's utterance. Here, there is a rise over 'alien' (perhaps L*+H), followed by a fall to the nuclear stressed syllable 'for-ces', where the final rise begins. The second plot represents an L1-transferring learner's utterance. Here, there is a high target on 'a-'lien), followed by an L1-like fall to the utterance-final syllable (for)'ces', where the final rise begins. Thus, the location of the final rise in these two utterances follows the general pattern sketched out above.

6.3.3 Results B: Frequency predicted by ratings

As was done above for the first two research questions, the potential role of proficiency (as indirectly indexed by the rating data) will now be discussed in its capacity as a predictor for
boundary rise alignment. The following plot follows the same format as those presented above for the first two research questions.

Figure 6.24: Rate of final syllable rising (y axis) as predicted by rating group

This time, unlike the previous two research questions, the rating data show a clear predictive role here. For all five tasks, tokens from the highly-rated group of learners show lower rates of final-syllable alignment than tokens from the low-rated group of learners. Interestingly, this comparison only reaches significance in a binomial test for the two segmental tasks: Sentence Segmentals (29/52=55.8% for Low vs. 28/73=38.4% for High) and Word Segmentals (26/48=54.2% for Low vs. 24/62=38.7% for High). However, a stairstepping trend is observed for all of the rating criteria except for Word Accent. Thus, it appears that the this particular variety of crosslinguistic intonation transfer differs from the other two discussed up to this point in that its rate of occurrence does appear to be proficiency-graded (at least to the extent that this can be concluded based on rating data).
It is technically possible that some low-proficiency learners rose at the final syllable of the utterance merely because that is when they saw the question mark (in their word-by-word processing of the sentence). However, this pattern was also observed in 38% of high-proficiency learners, for whom such processing constraints presumably have been overcome. This fact suggests that processing constraints are most likely not the only factor at work, and there is almost certainly some role for crosslinguistic transfer.\(^6\)

6.3.4 Revisiting Research Question 3

As outlined in Chapter 4 (§4.2.3), Research Question 3 is as follows:

(6) **Research Question 3:** In their L2 English production of interrogative sentences ending in one or more unstressed syllables, how often do Japanese EFL learners produce a simple boundary rise from within the final syllable?

The query used for this research question involved first identifying and setting aside all tokens lacking a final rise, then doing the same for double-rise contours. As such, the analysis was confined specifically to final fall-rise contours (e.g. a H* L* H- H% contour in English, or a H*+L L% in Japanese), which accounted for roughly 75% of tokens with final rises in learners and 50% of those in natives (cf. Table 6.12). For this specific kind of contour, rising from within the final syllable was markedly more common for learners than native speakers (cf. 46.6% vs. 11.9% in Table 6.13).

\(^6\) Since a question mark in English is visually grouped with the preceding word as a whole, even if a low-proficiency learner read a sentence word by word, they presumably would have seen the question mark by the point where they read the utterance-final word. As such, even for low-proficiency learners, it would not be expected for them to wait until the final syllable to rise; if anything, they should have risen from the beginning of the word. Thus, the processing-based account does not make the correct predictions even for the low-proficiency learners.
The answer to the research question is complicated somewhat by the rating data component of the analysis. There, the results suggested that the extent to which this phenomenon occurs in Japanese EFL learners is tied to proficiency. More specifically, beginning learners exhibited final-syllable alignment more than advanced learners (in approximately 55% vs. 38% of tokens, respectively). Thus, it appears that, as a learner gains proficiency in various aspects of L2 phonology (and perhaps especially segmental aspects), this particular kind of crosslinguistic transfer decreases in frequency.

With all three of the research questions now fully addressed, the stage is set for making higher-level inferences based on these findings. This is the focus of the seventh and final chapter of the dissertation.
Chapter 7: Conclusion

The present dissertation had three components. In its theoretical component (Chapter 2), it sketched out a proposal for a typology describing not only the space of possible ways two languages' intonation systems can mismatch but also the implications of such differences for crosslinguistic transfer in L2 acquisition. The methodological component (Chapter 3) described a novel approach to the stylization of F0 contours that allows for the parametrization of contour shape, thus creating a valuable visual feedback loop in the stylization process (as to whether the model is a reasonable fit to the data). The empirical component (Chapters 4 through 6) generated three research questions based on contrastive analysis, then explored them by querying stylizations of F0 data from a subset of the large L2 speech corpus "ERJ".

Chapter 6 provided evidence that all three kinds of crosslinguistic transfer under examination do in fact occur in Japanese EFL learners. The higher-level issue, set out in Chapter 1 (§1.2.3) as the goal of the empirical component of the dissertation, of which of these three is most frequent has yet to be systematically addressed. Since all three research questions shared the common structure, "How often do Japanese EFL learners do ___", it is now possible to assemble all of the results from Chapter 6 together into a bigger picture, namely the overall hierarchy of frequency at which the three phenomena in question transfer (§7.1). This is then followed by a general discussion of the broader theoretical implications of the results in section
7.2. Section 7.3 then concludes the dissertation with a discussion of its limitations and future directions.

7.1 Assembling results into a hierarchy

The simplest way of comparing the frequency of occurrence across the three research questions is by examining comparable data for each presented in §6.1.2, §6.2.2, and §6.3.2. The most directly comparable analysis presented in each of those sections is the percentage, for each combination of sentence and speaker group (native vs. learner), that the phenomenon in question occurred. (For Research Question 2, this is the dichotomized data, i.e. the rate at which low boundary marking occurred one or more times per utterance.) The relevant data from Chapter 6 is reproduced in the following table. (Note that the row numbers in the first column are arbitrary and are provided for cross-referencing only; sentence 1 for one research question is not the same as sentence 1 for another research question.)
Table 7.1: Comparison of rates of transfer across all three research questions

In the final 'Total' row of the table, the overall percentages are 10%, 48%, and 46.6% for Research Questions 1 through 3, respectively. This information alone suggests that, of the three, phrasal H- insertion (Research Question 1) is least common of the three. In contrast, the other two phenomena - increased rates of low boundary marking (Research Question 2) and final-syllable alignment in boundary rises (Research Question 3) - are comparatively much more common.

Since 14 sentences were selected for each of the research questions, another way of making the results across all of the research questions directly comparable is by examining the median frequency of transfer across the 14 sentences for each question. This simply involves taking the percentage occurrence for each sentence in each research question as 14 data points to calculate a median over. Such an analysis not only factors out irrelevant variation within the
arbitrarily selected set of materials but also can be thought of as a more realistic measure (i.e., for any given sentence chosen at random, what is the median likelihood a learner will produce it with the transfer pattern). When looked at in this way, the hierarchy is the same - 6.8% for Research Question 1 versus 56.55% for Research Question 2 and 48.35% for Research Question 3 Thus, this alternate perspective reinforces the conclusion made above.

The analyses above were based on the raw proportions of learner tokens that satisfied the requirements of the relevant query. A third way of globally looking at the results is to calculate how much more frequently learners' tokens did so relative to native speakers. This is the focus of the following table:

<table>
<thead>
<tr>
<th>#</th>
<th>Research Question 1</th>
<th>Research Question 2</th>
<th>Research Question 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>12 - 0 = 12%</td>
<td>66.7 - 54.5 = 12.2%</td>
<td>43.8 - 27.3 = 16.5%</td>
</tr>
<tr>
<td>2</td>
<td>12 - 0 = 12%</td>
<td>21.7 - 36.4 = -14.7%</td>
<td>40.9 - 0 = 40.9%</td>
</tr>
<tr>
<td>3</td>
<td>16 - 9.1 = 6.9%</td>
<td>53.8 - 10 = 43.8%</td>
<td>52.9 - 33.3 = 19.6%</td>
</tr>
<tr>
<td>4</td>
<td>4 - 0 = 4%</td>
<td>69.2 - 9.1 = 60.1%</td>
<td>20 - 0 = 20%</td>
</tr>
<tr>
<td>5</td>
<td>0 - 9.1 = -9.1%</td>
<td>74.1 - 40 = 34.1%</td>
<td>58.8 - 50 = 8.8%</td>
</tr>
<tr>
<td>6</td>
<td>40 - 9.1 = 30.9%</td>
<td>66.7 - 40 = 26.7%</td>
<td>53.8 - 0 = 53.8%</td>
</tr>
<tr>
<td>7</td>
<td>0 - 9.1 = -9.1%</td>
<td>70.8 - 9.1 = 61.7%</td>
<td>25 - 0 = 25%</td>
</tr>
<tr>
<td>8</td>
<td>19 - 9.1 = 9.9%</td>
<td>52 - 27.3 = 24.7%</td>
<td>81.8 - 0 = 81.8%</td>
</tr>
<tr>
<td>9</td>
<td>4.5 - 0 = 4.5%</td>
<td>12 - 0 = 12%</td>
<td>37.5 - 18.2 = 19.3%</td>
</tr>
<tr>
<td>10</td>
<td>14.3 - 9.1 = 5.2%</td>
<td>50 - 0 = 50%</td>
<td>6.2 - 0 = 6.2%</td>
</tr>
<tr>
<td>11</td>
<td>9.1 - 0 = 9.1%</td>
<td>14.3 - 9.1 = 5.2%</td>
<td>30.8 - 0 = 30.8%</td>
</tr>
<tr>
<td>12</td>
<td>0 - 9.1 = -9.1%</td>
<td>59.3 - 30 = 29.3%</td>
<td>60 - 0 = 60%</td>
</tr>
<tr>
<td>13</td>
<td>4.5 - 0 = 4.5%</td>
<td>7.1 - 9.1 = -2%</td>
<td>88.9 - 0 = 88.9%</td>
</tr>
<tr>
<td>14</td>
<td>0 - 0 = 0%</td>
<td>64.3 - 9.1 = 55.2%</td>
<td>83.3 - 50 = 33.3%</td>
</tr>
<tr>
<td>Total:</td>
<td>10 - 4.6 = 5.4%</td>
<td>48 - 20 = 28%</td>
<td>46.6 - 11.9 = 34.7%</td>
</tr>
</tbody>
</table>

*Table 7.2: Differences between learner and native speaker proportions for each sentence.*

The three numbers in each cell represent [learner] - [native] = [difference]%: Any positive number represents the state of affairs where a greater percentage of the learner tokens
satisfied the query than the native speaker tokens. The larger the number, the bigger the difference is. The results here look similar to those presented above: Research Question 1 is much lower (5.4%) relative to Research Questions 2 and 3 (28% and 34.7%, respectively). In terms of medians across the 14 sentences for each research question, the percentages are similar; 4.85% vs. 28% and 27.9%.

### 7.2 General discussion

To summarize the discussion up to this point, it appears that, for the sample of 202 Japanese EFL learners represented in the ERJ, the hierarchy of frequency is as follows:

<table>
<thead>
<tr>
<th>RQ</th>
<th>Typology node</th>
<th>Phenomenon</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>[Position, Emptiness]</td>
<td>spurious insertion of initial H-</td>
<td>Low</td>
</tr>
<tr>
<td>2</td>
<td>[Density, Edge tones]</td>
<td>marking of pre-pausal boundaries with L%</td>
<td>High</td>
</tr>
<tr>
<td>3</td>
<td>[Realization, Alignment]</td>
<td>beginning boundary rises in final syllable</td>
<td>High</td>
</tr>
</tbody>
</table>

Table 7.3: Summary of empirical results from the present dissertation

These results beg the question of why the phenomena for Research Questions 2 and 3 were especially frequent. Based on the contemporary understanding of L2 segmental phonology, it would clearly be premature to conclude that the present study's results fall out of an 'inherent' stratification of the typology nodes in terms of frequency, i.e. [Density, Edge tones] ≈ [Realization, Alignment] > [Position, Emptiness]. Indeed, it seems likely that there is most likely no strict universal ranking in terms of frequency among the 12 nodes in the typology from Chapter 2. Thus, rather than pursue the idea of a universal hierarchy, the overall goal for this avenue of research should be to identify what these various factors are as well as exactly how they impact the frequency of a given transfer phenomenon for a given L1-L2 pairing.
For Research Question 2, one attractive explanation is that low boundary marking itself is in fact fully targetlike, since native English speakers also frequently produce English L-. The relevant transfer phenomenon is only a matter of degree. As such, learners’ input is inherently ambiguous, such that a low F0 target before a prosodic boundary can be parsed as an L2 tone equally as well as an L1 tone. Late boundary rise alignment (Research Question 3) may have possibly been so frequent because its typology node, [Realization, Alignment], is inherently pervasive. It is well documented in the L2 segmental phonology literature that differences in the fine phonetic details between otherwise equivalent phonemes in two languages frequently causes acquisitional problems for L2 learners. Not surprisingly, recall from Chapter 2 (§2.4.1) that [Realization, Alignment] transfer has been reported in the literature in numerous studies on a wide variety of L1-L2 pairings.

Of course, it is difficult to fully rule out the possibility that methodological artifacts are partially responsible for the results. As acknowledged in Chapter 6, phrasal H- insertion (Research Question 1) may seem to be so rare simply because the phrasal H- appeared merely as an elbow in the contour rather than a clearly-defined target (and therefore did not receive a vertex). It could also have been so rare because some bona fide instances of phrasal H- were unduly discarded in the process of factoring out potential (L+)H* on the secondary stress. Along similar lines, low boundary marking (Research Question 2) could have appeared in such a high proportion of tokens simply because there were multiple chances for the relevant phenomenon to appear in a given sentence. Finally, the rates for late boundary rise alignment (Research Question 3) may in fact be slightly inflated due to the fact that non-rising and double-rising contours were
pre-emptively excluded. (Note that similar adjustments were not made for the other two research questions.)

Interestingly, all three kinds of transfer under investigation were observed at all proficiency levels (as indexed by the rating data). Only the nontargetlikeness in boundary rise alignment for Research Question 3 appeared to decrease with proficiency, and even there, it was only reduced to approximately 38% in highly-rated learners. Since all of the learners in the corpus were learners of English as a foreign language, it is possible that the other two kinds of transfer would also be proficiency-graded if more advanced learners (e.g. with residential experience abroad) were tested. It is also possible that if sufficiently advanced learners were found, the rates of transfer would approximate 0% much more closely. At least for the present dataset, however, the rating data overall predicted relatively little, and the hierarchy of frequency was more or less fixed across proficiency levels. In terms of a longitudinal trajectory, it therefore appears all three of the phenomena under investigation appear in beginning learners' speech and remain for quite some time.

Finally, it is worth noting that, while the present study was framed from the perspective of crosslinguistic transfer, other universal/developmental factors almost certainly also play a role. In other words, it is unlikely that the empirical results of the present dissertation are due to transfer alone. While the role of fluency and processing constraints was discounted in Chapter 6 as the primary determining factor in the results, they may still play a more minor role, perhaps reinforcing the transfer-based trends. A fuller understanding of this issue is only possible through future studies using materials designed explicitly for the purpose of teasing apart the role of transfer from universal/developmental factors.
7.3 Limitations and future directions

The present dissertation is only the first step in a larger research program. The three end-products of the dissertation - namely the typology, the method, and the hierarchy - all lend themselves to numerous promising directions for future research. Each of these can be thought of as stemming from some limitation of the present study. The purpose of the present section is to outline these limitations and future directions. Due to their large total number, and since each of them could be the subject of its own in-depth explanation, only a brief (1-2 sentence) outline of each future direction will be sketched out.

The following are the future directions for the theoretical component of the present dissertation (i.e. the typology in Chapter 2):

1. **Non-transfer factors**: Whereas the typology focused narrowly on crosslinguistic transfer, it would be beneficial to pursue other alternative typologies that are not so narrowly focused on this one particular aspect of L2 development and are more agnostic to the underlying causes of observed nontargetlikeness.

2. **Refining/expanding**: It is possible some kinds of transfer besides those discussed in Chapter 2 would be difficult to classify under the existing typology. Thus, existing nodes could be split (or new ones entirely added) in order to make it easier to apply the typology in practice.

3. **Perception/interpretation**: The present typology is focused narrowly on L2 production specifically. In the future, the existing typology could be either expanded to accommodate L2 intonation perception/interpretation or else a separate but parallelly-structured perception/interpretation typology could be constructed.
The following are the future directions for the methodological component of the present dissertation (i.e. the stylization framework sketched out in Chapter 3):

1. **Reliability/consistency**: Beyond presenting the distribution of weighted mean absolute deviation (MAD) values, there was no indication of how reliable or consistent the stylizations used in the present dissertation were. In future, having multiple people annotate the same file set, as well as having the same person annotate the same file multiple times, could serve as the foundation for a more systematic analysis of this issue.

2. **Automation**: All files were stylized manually for the present dissertation, which not only greatly increases human labor but also creates the risk for subjectivity. A natural next step would be to use the manual annotations from the present dissertation as a gold standard in an attempt to automate the stylization process while still keeping it phonologically meaningful.

3. **Break indices**: The present approach to stylization modeled the F0 contour as a series of vertices and transitions, which misses the break index component of ToBI entirely. In future, additional layers to the stylization can be added to take into account varying degrees of prosodic boundary strength.

4. **'Seed' clicking**: In stylization, the vertices were determined by having the annotator click on a single point they felt reflected their mental analysis of the contour. In future, it would be fruitful to make it possible to click on two points indicating the range of time where a vertex is likely, thus allowing the annotator to express their degree of uncertainty.
5. **Pre-emptive visualization**: At present, the annotator does not see the consequences of their choice of a vertex until after they have clicked, at which point they may need to use the 'undo' feature. In the future, it would be useful to explore the possibility of somehow showing the user (e.g. using coloration) which candidate vertices would be better/worse choices *before* they make their choice.

6. **Classification**: It is possible to build a "classification model" to take stylizations and automatically integrate them with a segmentation to generate tone labels (in, e.g., RaP or ToBI) based on a computationally-implemented model of the intonational phonology of a given language. While mildly intractable for L2 learner data, this is perfectly manageable task for native-speaker data (and indeed the present author has already developed prototypes for such algorithms).

The future directions for the empirical component of the present dissertation are especially numerous:

1. **More nodes**: The present dissertation only investigated three of the 12 total nodes in the typology. Future research can follow the same general approach pursued here but with the other 9 nodes, thus fleshing out the idea of a hierarchy of transfer frequency even further (most notably, with the inclusion of kinds of transfer relating to pragmatic/discourse function, not just phonetic/phonological form).

2. **Opposite directionality**: Together with the typology from Chapter 2, the empirical results of the present dissertation make clear predictions about what sorts of patterns are expected in the speech of L1 English learners of L2 Japanese. A future study on such a
population using maximally similar materials and methods as those applied here would provide a valuable counter-point to fill in the other half of the picture.

3. **Other L1-L2 pairings**: At present, it may not be safe to generalize the results of the present dissertation beyond the one specific L1-L2 pairing under examination. What is now needed is for future studies to test analogous language pairings and see if the same kinds of results found here are obtained there as well.

4. **Replication**: Inevitably, some aspects of the present dissertation's results are a byproduct of the specific choice of materials used (e.g. inflated rates for Research Question 2 due to its use of longer sentences with multiple possible boundary locations). If a future study using entirely different materials replicates the present study's findings, its conclusions would be put on more solid empirical grounds.

5. **Within-learner analysis**: The list structure of the ERJ precluded an analysis of the propensities for the different kinds of transfer within a single learner, hence all analyses had to be presented in the aggregate. Future laboratory studies can be designed to get at this issue more directly by having fully parallel materials for all learners.

6. **Longitudinal data**: Since the ERJ is a cross-sectional corpus, all analyses presented in the present dataset were based on a single point in time for each learner. In future, it would be revealing to have a learner come in at multiple points in time and have them perform the same battery of tasks to track longitudinal progress over time.

7. **Proficiency**: In the present dissertation, learners' proficiencies were surmised based on indirect evidence from how samples of their speech were evaluated in a rating task. In
future work, it would be ideal to obtain more direct measures of proficiency for each learner - both for phonology as well as L2 knowledge in general.

8. **Benchmarking queries**: In order to verify whether the queries used in Chapter 6 are properly structured and functioning as intended, it is important to benchmark them with a large-scale corpus of spoken English. Since the queries were designed to tap patterns transferred from L1 Japanese, the prediction is that they would return relatively few results in such an analysis.

9. **Deeper analysis of queries**: The hierarchy discussed in §7.1 may depend on how strict or lenient the queries from Chapter 6 were. Future work can more systematically test the consequences of each step in the query design process (e.g. the choice to filter out H+!H* for RQ1 or L* for RQ2).

10. **Fluency/processing**: As discussed in Chapter 6, fluency and processing constraints may exert an influence on the data. Future studies can shed light on this issue by examining the extent to which measures of fluency (such as pausing rate/duration or utterance length) predict learners’ outcomes for the three research questions.

11. **L2-to-L1 influence**: The present study examined only the influence of the L1 on the L2. Using a similar population (Japanese EFL learners), an examination into whether aspects of English intonation transfer back into L1 Japanese would be a fascinating next step for this strand of research.

12. **In-depth rating analysis**: The present dissertation barely scratched the surface of the rich rating data included in the ERJ. A future study ought to take up the intonation subset
of the rated sentences and investigate what specific aspects of the F0 contour have the strongest impact on raters' scores.

13. **Loanword influence**: Research Question 3 (and, to a lesser extent, Research Question 1) touched on issues relating to learners' lexical representations for specific words in L2 English, some of which may have loanword equivalents in the L1. Materials used in future studies would benefit from more systematically controlling the influence of such loanwords (or else use an L1-L2 pairing where such borrowing does not exist).

14. **Word frequency**: Since word frequency was not controlled for in the present study, a subset of the stimuli contained words learners may not have been familiar with (hence, e.g., a word's primary-stressed syllable in the learner's interlanguage may have been different than expected). A future study could (1) explicitly test this effect by comparing materials with high vs. low-frequency words, (2) restrict the materials to entirely high-frequency words, or (3) empirically test lexical knowledge (e.g. stress location) and/or word familiarity via learner self-report.

15. **Register/task**: The results of the present dissertation are based entirely off of pre-scripted speech from a read-aloud task conducted in a laboratory. It still remains to be seen whether the same results would be obtained for numerous other conceivable tasks and speech registers (e.g. spontaneous conversation).

16. **Positive transfer**: Based on differences between the two languages in question (English and Japanese), the present study took the *negative* (i.e. detrimental) transfer as its starting point. It would be fruitful to examine the converse case, i.e. empirically testing whether
similarities between these two languages lead to accelerated rates of targetlikeness (positive transfer).

17. **Perception/interpretation**: The present study was restricted entirely to examining Japanese EFL learners' intonation production. It is left to future research to determine whether the perceptual implications of the present dissertation's three research questions are in fact substantiated.

18. **Individual differences**: Since demographic information on the 202 learners was not included with the public release of the corpus, it was impossible to account for factors such as age of onset/acquisition or length of residence in L2-speaking locales. A complete answer to the research questions asked in the present dissertation is incomplete without taking into account such factors (as well as cognitive individual difference variables such as working memory).

As attested by such a lengthy list of future directions, a wide expanse of terrain has yet to be explored in the field of L2 intonation acquisition. It is hoped that, when taken together as a whole, the three components of the present dissertation have made some headway in charting out this territory.
References


Breen, M., Dilley, L. C., Kraemer, J., & Gibson, E. (2012). Inter-transcriber agreement for two systems of prosodic annotation: ToBI (Tones and Break Indices) and RaP (Rhythm and Pitch). *Corpus Linguistics and Linguistic Theory, 8*(2), 277-312.


International Conference on English Language and Anglophone Literatures Today (pp. 238-249). Novi Sad, Serbia: University of Novi Sad.


Curriculum Vitae

EDUCATION

- **Indiana University – Bloomington:** *Fall 2008 - Summer 2015*
  - Ph.D., Linguistics
  - Ph.D., Second Language Studies
- **University of Nebraska Omaha:** *Fall 2004 - Spring 2008*
  - B.A., English (Writing and Linguistics Concentration)
  - B.A., Spanish
- **Creighton University:** *Fall 2002 - Spring 2004*
  - Non-degree

PUBLICATIONS


Assistantships

Research

- Yoshihisa Kitagawa (Linguistics): Spring 2010 - Summer 2011
- Charles Lin (East Asian Languages and Cultures): Summer 2012-Summer 2013

Teaching

- Instructor, Introduction to the Study of Language: Summer 2012
- Associate Instructor, Introduction to Linguistic Analysis: Fall 2011
- Associate Instructor, First-Year Japanese: Fall 2008 - Fall 2009, Fall 2013

Honors and Awards

Indiana University

- Dissertation Completion Fellowship: Fall 2014 - Summer 2015
- Householder Paper Award: Spring 2011, Spring 2012
- College of Arts and Sciences Travel Award: Spring 2015
- East Asian Studies Center Travel Grant: Fall 2013, Spring 2015
- Linguistics Department Householder Travel Grant: Fall 2011
- Pan-Asia Institute Travel Grant: Spring 2011

University of Nebraska Omaha

- Regent’s Scholarship: Fall 2004 - Spring 2008
- Hansen Distinguished Scholarship: Fall 2004 - Spring 2008

Service

Reviewer

- Acoustical Society of America: Summer 2014
- Indiana University Linguistics Club Working Papers Online: Summer 2010

Government

- Indiana University Linguistics Club: Secretary 2010-2011, President 2011-2012
- Second Language Studies Graduate Student Organization: Secretary 2010-2011

Professional Affiliations:

Linguistic Society of America (2008-2015)