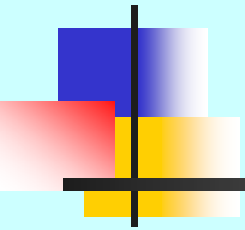


International Workshop on Grammar and Evidence, April 14-15, 2007

Corpus-based Research on Child Phonology



Jane Tsay

Institute of Linguistics

National Chung Cheng University, Taiwan



Outline

- Background: Theoretical issues
- Methodology: A corpus-based approach
- Testing domains
 - Syllable acquisition
 - Tone acquisition



Grammar and the Lexicon

- Grammar: Universal innate markedness
- Lexicon: Language specific lexical properties



Grammar has been the focus

- In the literature of child language acquisition, researchers have focused primarily on universal innate patterns, described in current phonological theories with markedness constraints.



Markedness

- Universal innate tendencies or patterns (since Prague School, e.g. Jakobson 1941/68, Trubetzkoy 1958/69)
- feature, feature value, rules, representation



Characteristics

- Unmarked: natural, innate, universal, easier to produce or perceive, formally simpler, less formal cost
- Marked: unnatural, learned, language specific, more formal cost



Markedness Constraints in OT

- Optimality-theoretic (OT) models of child language acquisition predict:
 - In the earlier stage: Markedness/Structure constraints >> Faithfulness constraints
 - Constraint reranking in the later stage
- Prince & Smolensky 1997; Tesar & Smolensky 1998; Boersma & Hayes 2001



Frequency Information in the Lexicon

- Sound patterns found in the adult lexicon which contains crucial information about frequency, including both type frequency and token frequency
 - Menn & Stoel-Gammon 1995; Tyler & Edwards 1998; Gierut, Morrisette, & Champion 1999; Bybee 2001



Methodology

- Child language data
- Corpus-based analysis: e.g. syllable frequency counts



Taiwan Child Language Corpus

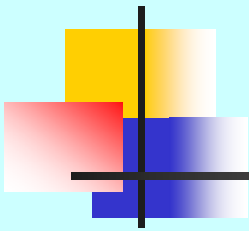
台灣兒童語料庫 (Tsay, in preparation)

- Longitudinal data (about 330 hours)
- Conversations between children and adults in Taiwan Southern Min speaking families



Data Collection

- Target language: Southern Min Chinese spoken in Taiwan (also called Minnan, Taiwanese, Amoy in the literature)
- Children: 9 boys, 5 girls
- Age range: approximately 1;6 to 4;0
- Recording: regular home visits (every two to three weeks); 40-60 minutes per session



Name	Sex	Age	Sessions	Duration (minutes)
YDA	M	3;11 – 4;4	9	540
YCX	M	3;10 – 4;0	6	285
LJX	M	3;9 – 4;2	8	530
CQM	M	2;9 – 4;6	30	1584
LMC	F	2;8 – 5;3	50	2045
YJK	M	2;6	2	105
CEY	F	2;1 – 3;10	37	1728
HBL	M	2;1 – 4;0	45	1889
LWJ	F	2;1 – 3;7	36	1777
WZX	M	2;1 – 4;3	44	1757
YSW	M	1;7 – 2;7	21	1210
TWX	F	1;5 – 3;6	44	1829
HYS	M	1;2 – 3;4	51	2280
LYC	F	1;2 – 3;3	48	2255
Total			431	330 hours



Corpus size

- 431 recording sessions (40-60 minutes per session)
 - 431 text files
 - 431 sound files



Corpus size

	Lines (utterances)	Words	Syllables	
			syllables (in content words)	syllables (in function words/particles)
			1,558,408	538,992
Total	497,426	1,646,503	2,097,400	



Transcription

- Orthographic transcription
 - 1. Romanization (Southern Min Pinyin)
 - 2. Chinese characters
- Phonetic transcription
 - 3. Segments (in Unicode IPA)
 - 4. Tones (in digits)



Transcription

- Pinyin: kan1 cit8 pue1
 'drink-up one cup (cheers)'
- Chinese: 乾 一 杯
- Segments: k a m t i t p u e
- Tone: 55 3 21



CHILDES (CHAT format)

Child Language Data Exchange System
(CHILDES, MacWhinney 1995)

- Headers
 - obligatory headers
 - constant headers
 - changeable headers
- Tiers
 - main tiers
 - dependent tiers



Main text

Main tier

*CHI: kan1 cit8 pue1.

Dependent tiers

%ort: 乾 一 杯.

%pho: k a m t i t p u e

%ton: 55 3 21



Annotation: syntactic categories

Main tier

*CHI: kan1 cit8 pue1.

Dependent tiers

%ort: 乾 一 杯.

%cod: VH Neu Nf

%pho: k a m t i t p u e

%ton: 55 3 21



Testing domain 1: the syllable

- All languages have syllables.
- The syllable is a fundamental phonological unit.
- In phonological theories, CV has been claimed to be the **core** syllable
 - Onset Principle: V.CV *VC.V
 - ONSET constraint (in Optimality Theory)



The syllable

- Infant vocalization development in Mandarin acquiring infants:
CV occurs first (Chen 1999; Chen & Kent 2005)



Q1: Markedness and frequency

- Is the most unmarked syllable type CV the most frequent syllable in child language?



Q2: The role of input in acquisition

- Is there a correlation between child speech and adult speech regarding syllable frequency?



Q3: Frequency and accuracy

- Is there a correlation between syllable frequency and accuracy?
(Are more frequently occurring syllable types acquired earlier?)



Four steps in counting syllable frequencies

- 1. Segment all syllables in the transcripts (in the main tier)
- 2. Count syllables
- 3. Code syllable types
 - sa → CV
 - ka → CV
- 4. Count frequency of each syllable type



Syllable segmentation

- Replace the digits for tone categories with a space
 - In our notation system, each syllable ends with a number for its tone category, e.g., *au7piah4* 後壁 "behind".
 - This notation system automatically provides the syllable boundary.



Count syllables (token frequencies)

- CLAN has a program `FREQ` that counts word frequencies.
- After all syllables are segmented by space, they are treated by `FREQ` as individual words.

au7piah4 → au piah → CV CVVC

A sample of the output of FREQ

```
Clan - [CEY2;l-sylseg.frq.cex]
File Edit View Tiers Mode Window Help
|freq +u +f +t *CHI @
Fri Sep 02 11:46:40 2005
freq (23-Sep-2003) is conducting analyses on:
  ONLY speaker main tiers matching: *CHI;
*****
From file <d:\freq-mark\research\syllable-types\syl-seg\cey-checked\sylseg\CEY01-sylseg.cha> to file <D:\Freq-Mark\research\syllabl
1 -
3 a
50 a0
3 ai
2 ai0ia0
3 ai0io0
1 an
1 bak
5 beh
18 bo
2 cai
66 ce
1 cha
2 chah
2 chiu
2 choo
2309[E][TEXT] *1
Ready
NUM
上午 10:07
星期二
2006/10/24
```

Results: Syllable frequencies

	Adults	%	Rank	Children	%	Rank
CV	382,760	33.2	1	140,028	34.5	1
CVC	260,358	22.6	2	79,976	19.7	2
CVV	209,672	18.2	3	79,763	19.7	3
V	122,111	10.6	4	47,426	11.7	4
CVVC	71,852	6.2	5	20,092	5.0	5
VC	28,341	2.4	6	8,438	2.1	6
VV	26,126	2.3	7	8,389	2.1	7
CN	21,392	1.9	8	7,661	1.9	8
N	12,293	1.1	9	5,812	1.4	9
CVVV	8,723	0.8	10	3,563	0.9	11
VVC	8,655	0.8	11	4,278	1.1	10
VVV	490	0.0	12	209	0.1	12
Total	1,152,773			405,635		



Syllable token frequencies: top two

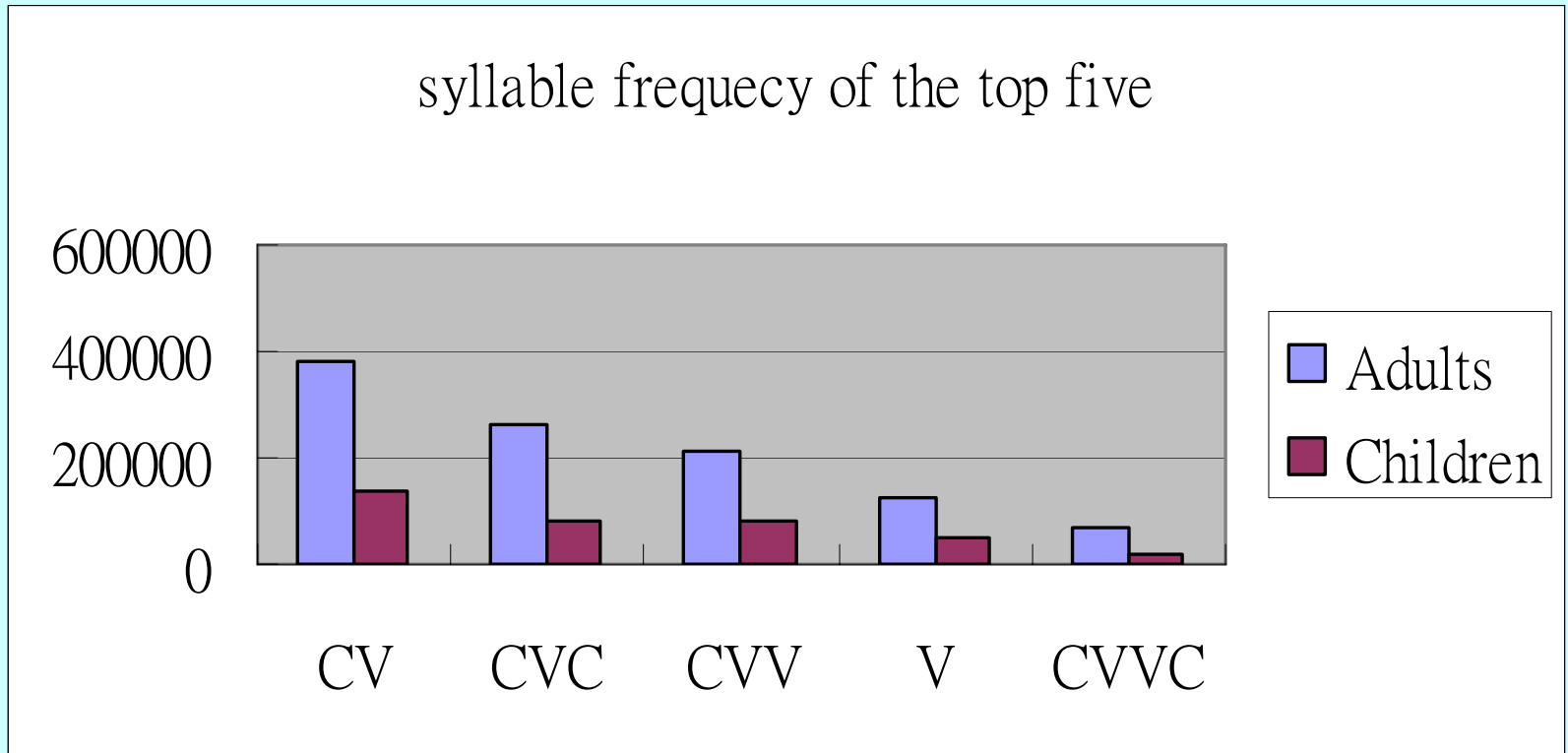
	Adults	%	Rank	Children	%	Rank
CV	382,760	33.2	1	140,028	34.5	1
CVC	260,358	22.6	2	79,976	19.7	2
Subtotal	643,118	55.7		220,004	54.2	



Syllable token frequencies: top five

	Adults	%	Rank	Children	%	Rank
CV	382,760	33.2	1	140,028	34.5	1
CVC	260,358	22.6	2	79,976	19.7	2
CVV	209,672	18.2	3	79,763	19.7	3
V	122,111	10.6	4	47,426	11.7	4
CVVC	71,852	6.2	5	20,092	5.0	5
Subtotal	1,046,753	90.8		367,285	90.5	

Syllable frequency





Summary

1. CV is the most frequent syllable and CVC the second most frequent

- $CV + CVC > 50\%$
- $CV + CVC + CVV + V + CVVC > 90\%$

2. Children and adults have similar patterns regarding token frequencies:

$CV > CVC > CVV > V > CVVC$



Frequency and accuracy

- More frequent → more accurate?



TWX 1;7 Syllable Errors

	Correct	Wrong	Total	Error rate (%)
CV	630	387	1016	38
V	554	54	608	9
CVV	164	188	352	53
CVC	7	319	326	98
CVVC	2	113	115	98
VC	1	0	1	0
VV	38	29	67	43
CN	1	21	22	95
N	17	4	21	19
VVC	0	18	18	100
CVVV	3	19	22	86
VVV	0	0	0	0
Total	1417	1152	2568	45



TWX 1;7 Syllable Errors

- 1. Frequency does not always have a positive correlation with accuracy rate
 - more frequent → more accurate?
 - “Yes” for CV and V
 - “?” for CVV
 - “No” for CVC

Statistics?



TWX 1;7 Syllable Errors

- 2. Most errors involved a coda C
 - CVC (98% error rate)
 - CVVC (98% error rate)
 - VVC (100% error rate)
- Not involving a coda C
 - CN (syllabic N) (95% error rate)
 - CVVV (triphthong) (86% error rate)



Coda consonants in child language

- Coda dropping is very common in early phonological acquisition (King 1980, Su 1985, Hsu 1989, So and Dodd 1995, Tsay & Huang 1998)
- Physiological constraints in motor control



Optimality Theory's Predictions about Child Language Acquisition

- Prince and Smolensky (1993, 1996)
- Early stage:
Structure/markedness Constraints >>
Faithfulness Constraints
 - Onset: CV
 - NoCoda: *CVC
- Constraint re-ranking in a later stage?
 - (further research)



Discussion

- 1. The results were consistent with the markedness hypothesis:
CV is the most frequent syllable
How about other syllable types?



Discussion

- 2. The ranking of syllable token frequencies in children's speech

$CV > CVC > CVV > V > CVVC$

is identical with that of the adults' speech

- Influence from the input (adults' speech)
- Universal?



Discussion

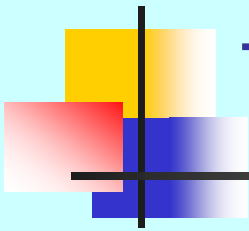
- 3. There doesn't seem to be a straightforward correlation between syllable token frequency and accuracy.



Discussion

4. Dutch syllable acquisition

- Levelt, Schiller, and Levelt (1999); Boersma & Levelt (1999)
- $CV \rightarrow CVC \rightarrow V \rightarrow VC \rightarrow$
- $\{CVCC \rightarrow VCC \rightarrow CCV\} \rightarrow \{CCVC \text{ or } CCV \rightarrow CCVC \rightarrow CVCC \rightarrow VCC\} \rightarrow$
- $CCVCC$



Testing domain 2: lexical tone

- Markedness in tone based on tone production (Ohala 1978; Maddieson 1978)
 - Rising requires more energy than falling
 - Lower pitch is more difficult to produce than higher pitch
 - M is the resting pitch (default?)



Marekedness in tone

- Based on tone production
 - $HL > LH$
 - $H > L$
 - M is the easiest?
 - LH is the most difficult being both rising and low register
- $M > \{HL, H\} > L > LH$



Q1: Markedness and frequency

- Is the most unmarked tone type M the most frequent syllable in child language?



Q2: The role of input in acquisition

- Is there a correlation between tone distribution in child speech and tone distribution in adult speech?



Q3: Frequency and accuracy

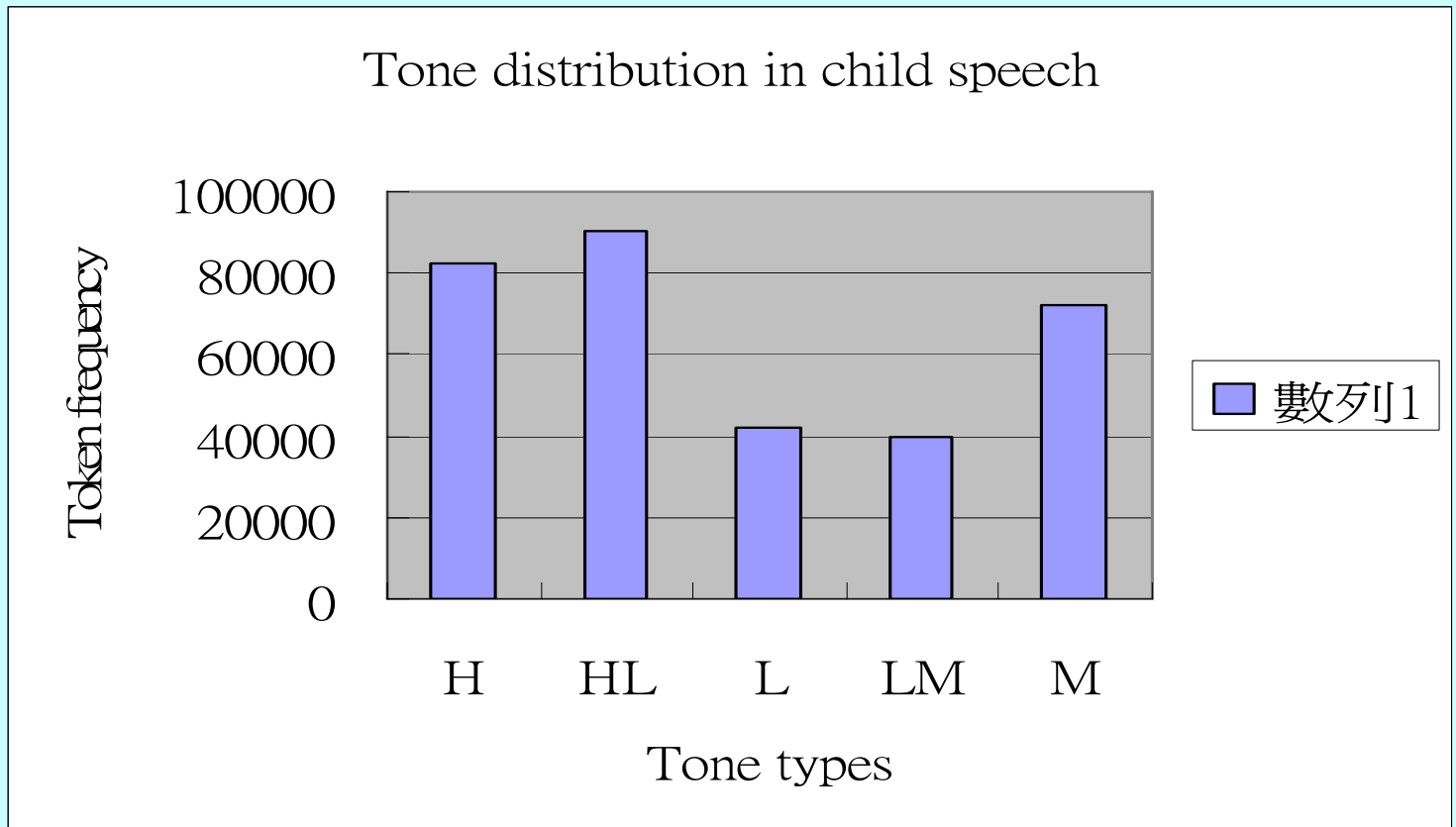
- Is there a correlation between tone type frequency and accuracy?
(Are more frequently occurring tones acquired earlier?)



Southern Min tone system

Tone category	Tone 1 陰平	Tone 2 陰上	Tone 3 陰去	Tone 4 陰入
tone type	high	falling	low	low-abrupt
Tone category	Tone 5 陽平		Tone 7 陽去	Tone 8 陽入
tone type	rising		mid	high-abrupt

Results (1)





Results (1): Markedness and frequency

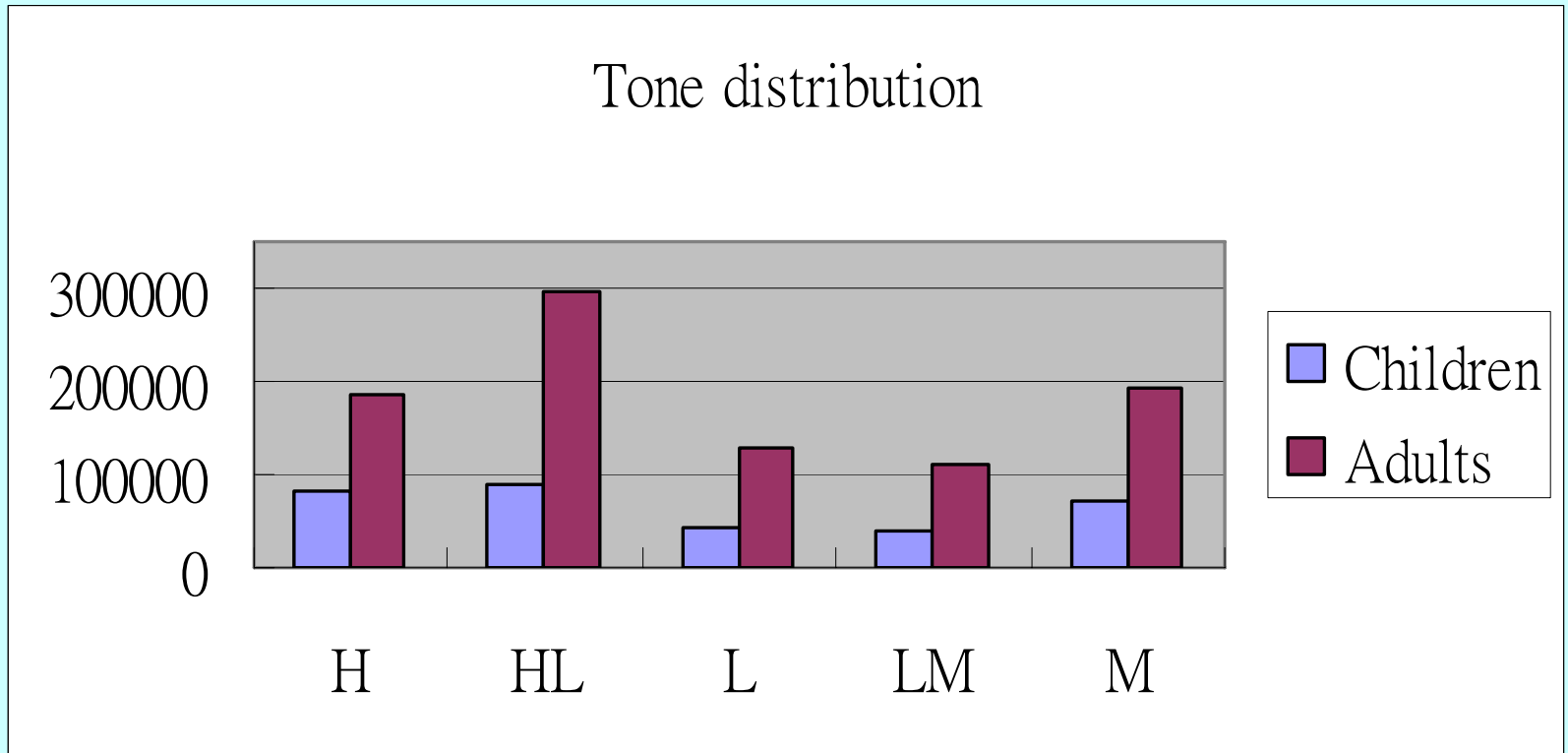
- Prediction: $M > \{HL, H\} > L > LH$
- Token frequency in child speech:
 - $HL > H > M > L > LM$
- HL is the most frequent tone in child speech
- M?? (physiology \neq linguistic system?)

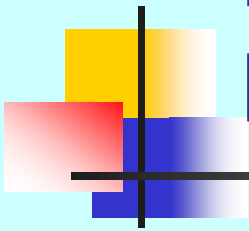


Results (2): the role of input

- Similar tone distribution between children and adult
- Token frequency in child speech:
 - $HL > H > M > L > LM$
- Token frequency in adult speech:
 - $HL > M > H > L > LM$
 - (M~H probably not significant)

Tone distribution

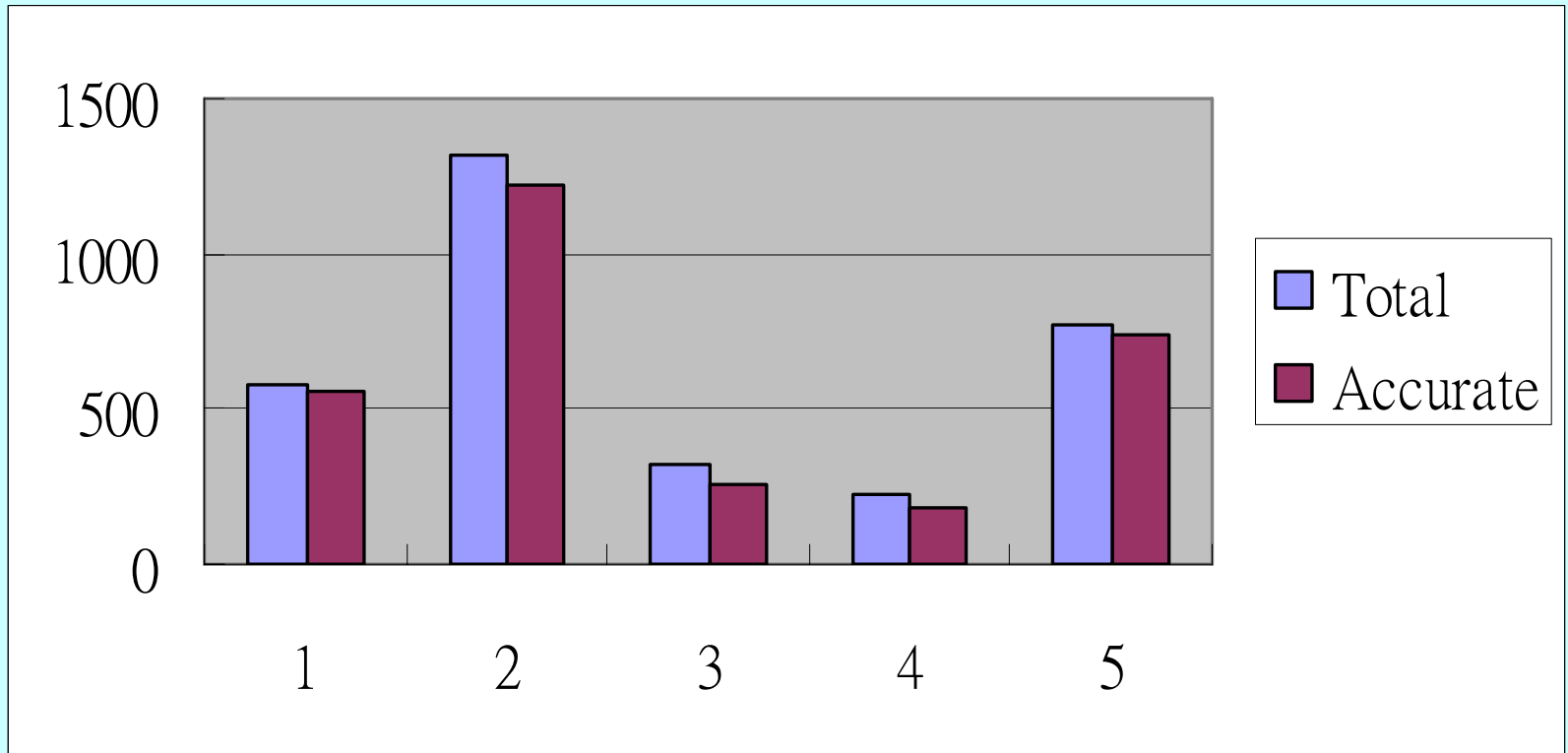




Results (3): Frequency and accuracy

- There is a positive correlation between tone frequency and accuracy

Tone accuracy rates (3 children 1;10-2;0)



Tone error analysis:

Grammar was playing a role

- Most of the errors in children's tone production were so-called "sandhi error" meaning that they mistakenly produced those tones in their corresponding sandhi form.
- This shows that children were not just imitating the adults and, more importantly, grammar (tone sandhi rules or OT constraints) was playing a role in children's tone production.



Summary

- 1. Markedness seemed to play a role, but not always
- 2. Similar patterns in tone distribution were found in children and adult speech
- 3. There was a positive correlation between tone frequency and accuracy
- 4. Error analysis showed that grammar was playing a role



Acknowledgements

- NSC grants
- Children and their parents
- RA's
 - Rose Huang, Kay Chen, Joyce Liu
 - Peggy Hsieh, Yunwei Li
 - Many others