Are Mandarin (& Hokkien)Tones Optimized for Efficient Communication?

2nd International Conference on Spoken Chinese Corpora, Aug 10 2016

Yuan-Lu Chen
cheny@email.arizona.edu

Background: information theory and language

Cross-linguistically the **length of a word** is predicted by its **frequency** and the amount of **information content** it has: the more frequent and with less information content the shorter a word is (Sigurd et. al. 2004, Piantadosi et al. 2011).

{more frequent, less information content} → short

{less frequent, more information content}→ long

Background: What is information content?

- Information content defined as how unexpected it is for a word to occur in a certain context.
- Zero information content: “United States of **America**”
- High information content: “The boy saw a **Unicorn**”

Background: information theory and language

{more frequent, less information content} → short

{less frequent, more information content}→ long

Human language lexical systems result from an optimization of communicative pressures (Piantadosi et al. 2011).

Figure adapted from Piantadosi et al. (2011:2): the correlations in English

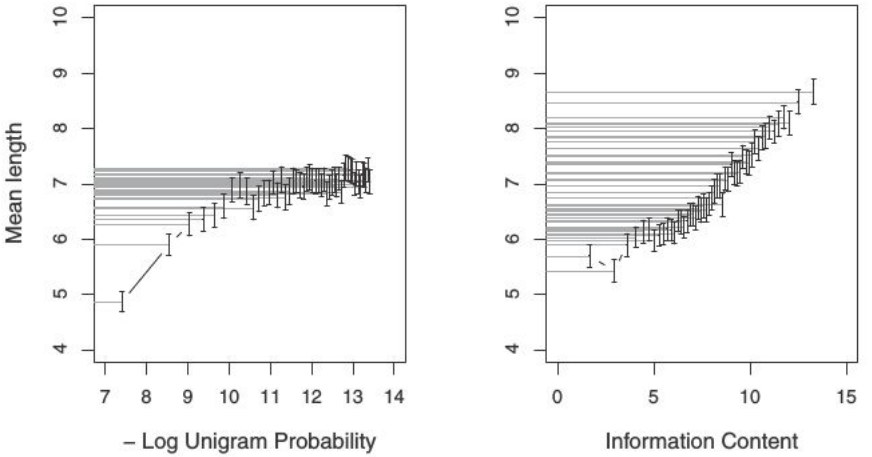


Fig. 2. Relationship between frequency (negative log unigram probability) and length, and information content and length. Error bars represent SEs and each bin represents 2% of the lexicon.

Background: information theory and language

However, Piantadosi et al. (2011) only look at Romance languages, in which length is a possible dimension of the shape of a word. In Mandarin, length is not a possible dimension, and instead tone is.

Background:

- There are four lexical tones in Mandarin:

Name	Tone Value	Description	Example
Tone1	55	high-level	ma1 ‘mother’
Tone2	35	high-rising	ma2 ‘hemp’
Tone3	214	low-dipping	ma3 ‘horse’
Tone4	51	high-falling	ma4 ‘scold’

- Tone1 (High-level) and Tone4 (falling) are acquired by infants before the Tone2 (rising) and Tone3 (dipping) (Li & Thompson 1977).
- Thus it is assumed that **Tone1 and Tone4 are simple** and **tone2 and tone3 are complex.**

Research Question

- How does the tonal system respond to the communicative pressures?
- How to define “length/cost” of tone?
- Are the complexities of tones optimized in a similar manner as word length?
  - {Tone1,Tone4} simple-> more frequent and less information content
  - {Tone 2, Tone3} complex-> less frequent and more information content

  (The answer is NO!)

Corpus Study (Mandarin)

- Corpus: Glenn et al. (2013) transcribe the news broadcast in China using Chinese characters. There are 1,536,150 characters (including punctuation).
- Processing the data:
  - **Original data**: 今天资深记者任永蔚独家访问卫生部长高强,医疗体制改革成焦点 话题.
  - **Converting the Chinese characters to tone-annotated pinyin using Pypinyin (2014, January 1):**
    - ji1n tia1n zi1 she1n ji4 zhe3 re4n yo3ng we4i du2 jia1 fa3ng we4n we4i she1ng bu4 cha2ng ga1o qia2ng , yi1 lia2o ti3 zhi4 ga3i ge2 che2ng jia1o dia3n hua4 ti2 .

Average info content

$$-\frac{1}{N}\sum_{i=1}^{N}\log P(W = w \mid C = c_i)$$ (Piantadosi et al. 2011)

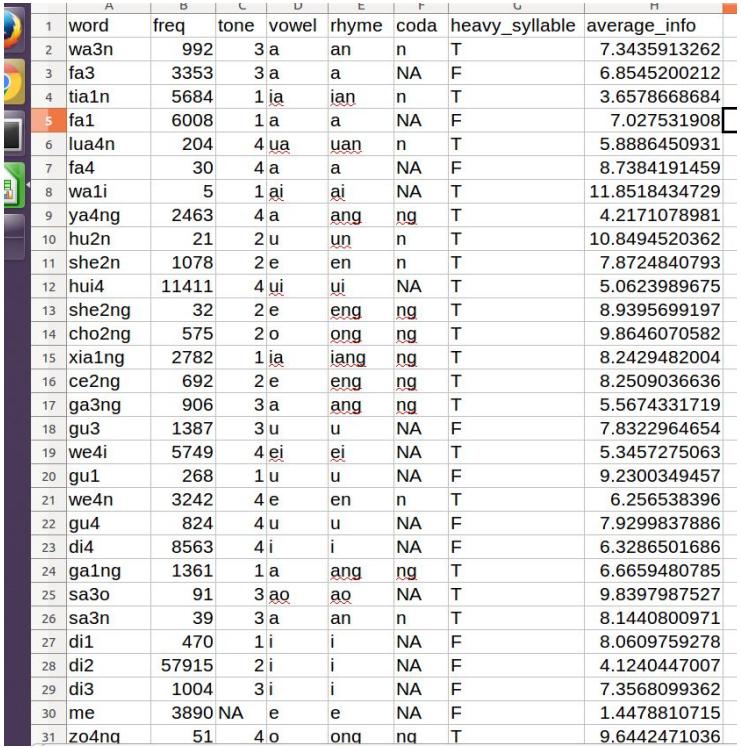
- e.g. ave_info(w_k)= -⅓ * (logP(w_k|c_1)+logP(w_k|c_2)+logP(w_k|c_3))
- -logP(w_k|c_i) ‘the unexpectedness of w_k preceded by c_i’
- P(w_k|c_i)=P(c_i_w_k)/P(c_i) = (freq(c_i_w_k)/sum_bi)/(freq(c_i)/sum_uni))
- sum_bi= sum of the bi-gram frequencies
- sum_uni= sum of the unigram frequencies

Average info content

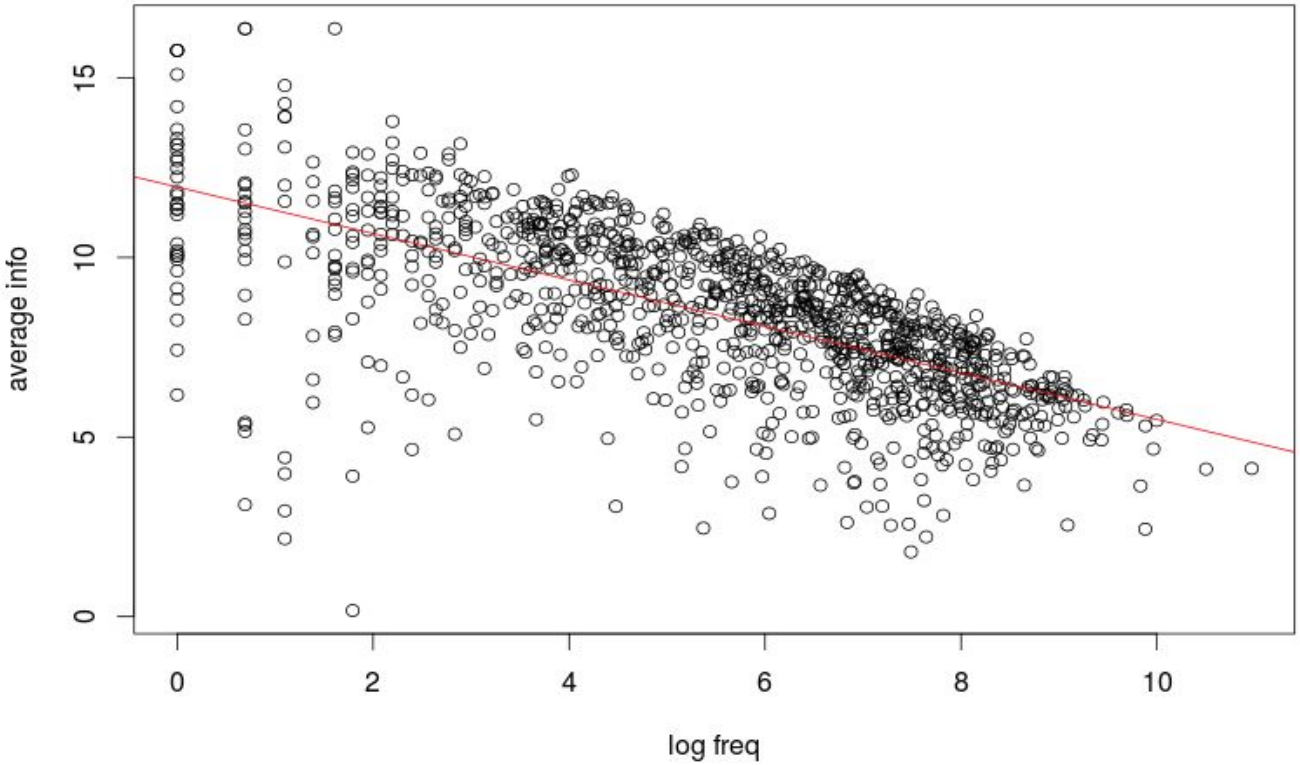
$$-\frac{1}{N}\sum_{i=1}^{N}\log P(W = w \mid C = c_i)$$

- ave_info(‘Tom’)
- = -⅓ * (logP(‘Tom’|‘saw’)+logP(‘Tom’|‘loves’)+logP(‘Tom’| ‘hates’))
- = -⅓ * (log(P(‘saw_Tom’)/P(‘saw’))+log(P(‘loves_Tom’)/P(‘loves’))+logP((‘hates_Tom’)/P(‘hates’)))
- = -⅓ * (log((2/100)/(10/101))+log((3/100)/(5/101))+log((4/100)/(4/101)))
- =1.0052
-
- $P(w_k|c_i)=P(c_i\_w_k)/P(c_i)$ = (freq($c_i\_w_k$)/sum_bi)/(freq(c_i)/sum_uni))
- freq(‘saw’)=10; freq(‘loves’)=5; freq(‘hates’)=4;
- sum of the unigram frequencies=101
- freq(‘saw_Tom’)=2; freq(‘loves_Tom’)=3; freq(‘hates_Tom’)=4;
- sum of the bi-gram frequencies=100
-

Data structure



	A	B	C	D	E	F	G	H
1	word	freq	tone	vowel	rhyme	coda	heavy_syllable	average_info
2	wa3n	992	3	a	an	n	T	7.3435913262
3	fa3	3353	3	a	a	NA	F	6.8545200212
4	tia1n	5684	1	ia	ian	n	T	3.6578668684
5	fa1	6008	1	a	a	NA	F	7.027531908
6	lua4n	204	4	ua	uan	n	T	5.8886450931
7	fa4	30	4	a	a	NA	F	8.7384191459
8	wa1i	5	1	ai	ai	NA	T	11.8518434729
9	ya4ng	2463	4	a	ang	ng	T	4.2171078981
10	hu2n	21	2	u	un	n	T	10.8494520362
11	she2n	1078	2	e	en	n	T	7.8724840793
12	hui4	11411	4	ui	ui	NA	T	5.0623989675
13	she2ng	32	2	e	eng	ng	T	8.9395699197
14	cho2ng	575	2	o	ong	ng	T	9.8646070582
15	xia1ng	2782	1	ia	iang	ng	T	8.2429482004
16	ce2ng	692	2	e	eng	ng	T	8.2509036636
17	ga3ng	906	3	a	ang	ng	T	5.5674331719
18	gu3	1387	3	u	u	NA	F	7.8322964654
19	we4i	5749	4	ei	ei	NA	T	5.3457275063
20	gu1	268	1	u	u	NA	F	9.2300349457
21	we4n	3242	4	e	en	n	T	6.256538396
22	gu4	824	4	u	u	NA	F	7.9299837886
23	di4	8563	4	i	i	NA	F	6.3286501686
24	ga1ng	1361	1	a	ang	ng	T	6.6659480785
25	sa3o	91	3	ao	ao	NA	T	9.8397987527
26	sa3n	39	3	a	an	n	T	8.1440800971
27	di1	470	1	i	i	NA	F	8.0609759278
28	di2	57915	2	i	i	NA	F	4.1240447007
29	di3	1004	3	i	i	NA	F	7.3568099362
30	me	3890	NA	e	e	NA	F	1.4478810715
31	zo4ng	51	4	o	ong	ng	T	9.6442471036

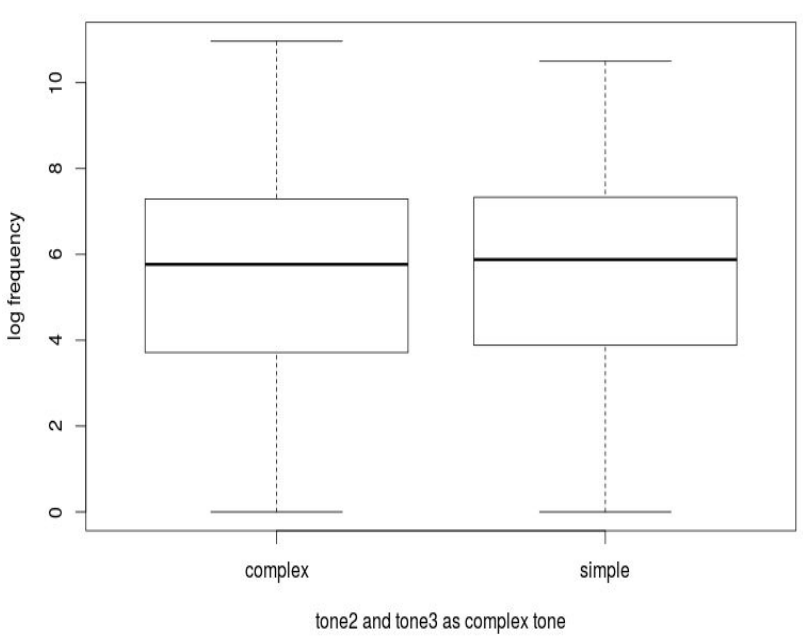


$R^2 = 0.4385$, $F(1,1128)=881$, $P<0.001$

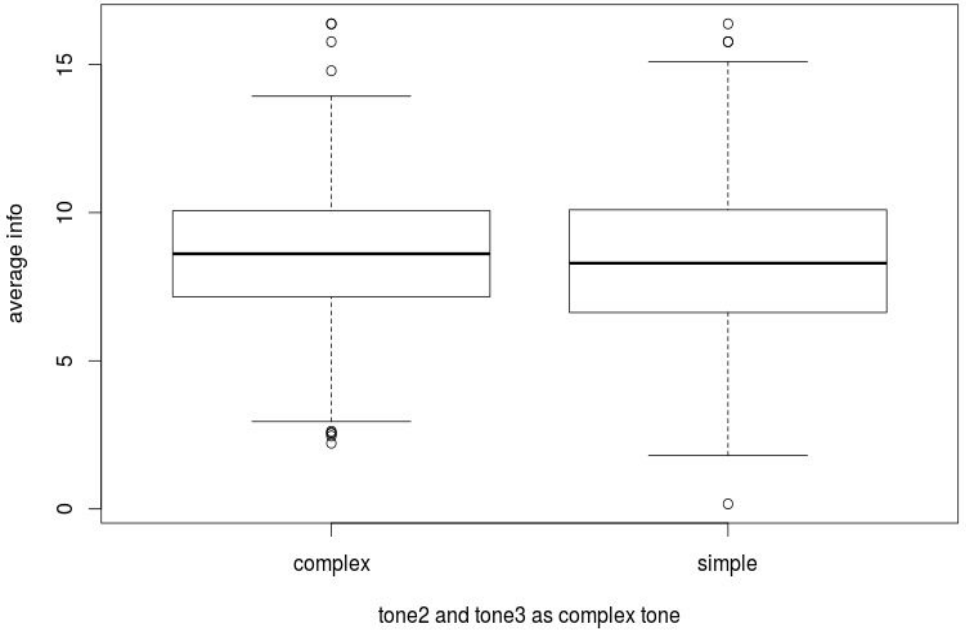
Result: freq~complex tone; average info~complex tone

- complex tone = Tone2 or Tone3
- Prediction: complex tone -> less frequent & more information content

log freq~complex tone
ANOVA (F(1,1128) = 1.397, p >.433)



average info~complex tone
ANOVA (F(1,1128) = 1.57, p >.12)



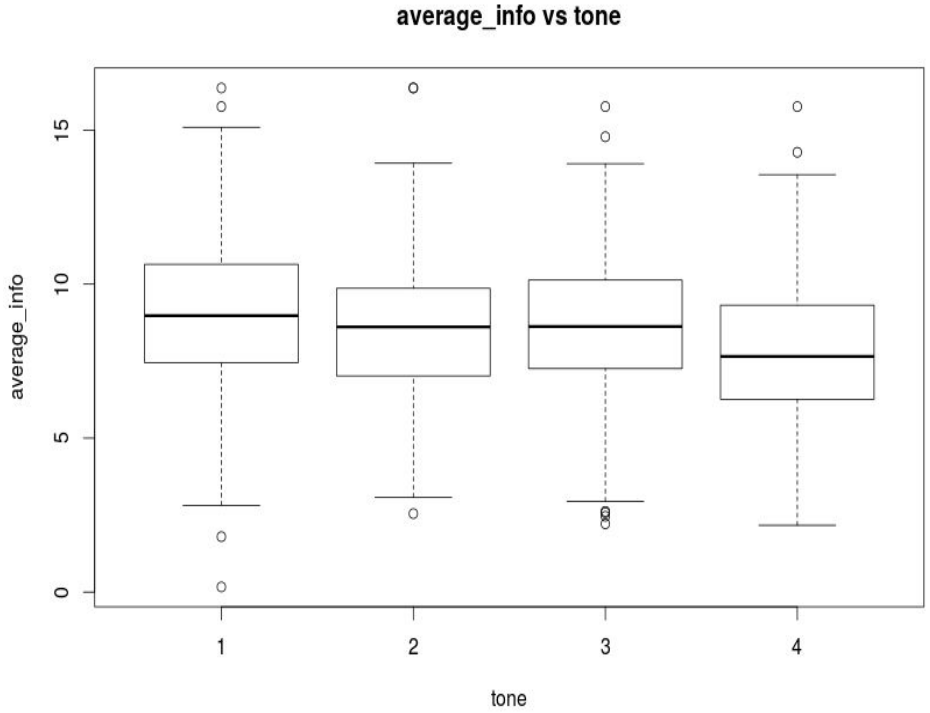
Result

- No evidence shows that the assumed complex tones (T2 and T3) have more information content or less frequent.
- However, tone matters!

Mandarin Result (average info)

average info~ tone, ANOVA (F(3,1126) = 13.78, p <.0001)

Tukey hsd post hoc test: 1>3>2>4

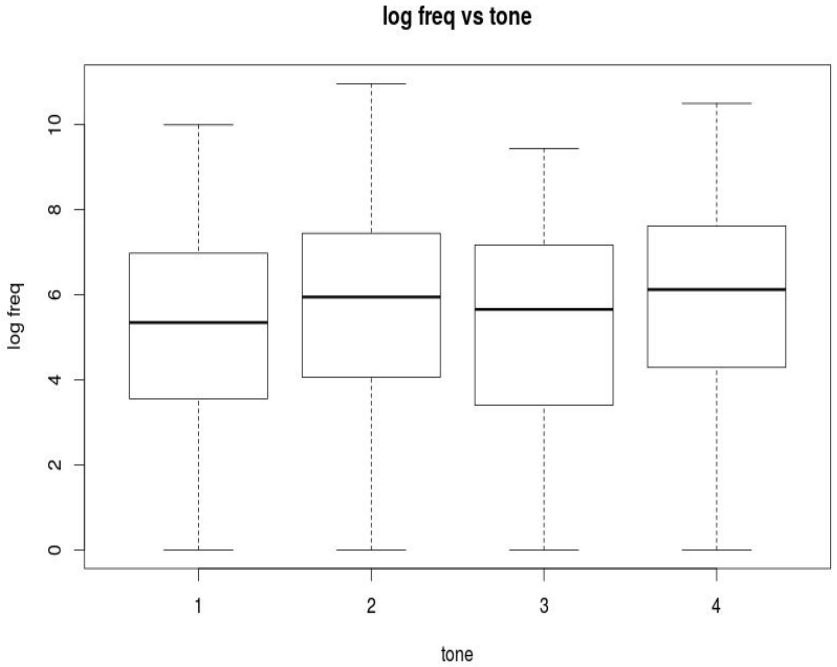


	diff	lwr	upr	p adj
2-1	-0.5155876	-1.0210448	-0.01013051	0.0435673
3-1	-0.3262441	-0.8119706	0.15948246	0.3094911
4-1	-1.1349859	-1.6037639	-0.66620792	0
3-2	0.1893436	-0.3249808	0.70366794	0.7793298
4-2	-0.6193983	-1.1177476	-0.1210489	0.0077575
4-3	-0.8087418	-1.2870675	-0.3304162	0.0000874

Mandarin Result (frequency)

log freq~ tone, ANOVA (F(3,1126) = 6.99, p <.001)

Tukey hsd post hoc test: 1<3<2<4



	diff	lwr	upr	p adi
2-1	0.51086708	-0.01093683	1.032671	0.0575631
3-1	0.04609228	-0.45534293	0.5475275	0.9953444
4-1	0.71964356	0.23570501	1.2035821	0.0007911
3-2	-0.4647748	-0.99573271	0.0661831	0.1101136
4-2	0.20877648	-0.30568979	0.7232427	0.7234925
4-3	0.67355128	0.1797563	1.1673463	0.0026272

Conclusion & Future Research

- T1 has more information content; T4 has least. T3 and T2 are in the middle.
- The assumed complex tones (i.e. T3 and T2) do not have more information content than the assumed simple tones (i.e. T1 and T4).
- The tones with most and least information (T1 and T4) are acquired first.
- A independent way to measure the cost/complexity is needed.
- How to incorporate the effect of T3 sandhi?
- Research on other tone languages: Taiwanese, Cantonese ... (I am looking for good corpus with tone annotated).

Taiwanese/Southern Min/Hokkien

Hokkien corpus

- 蔡素娟, 麥傑. 2013. 臺灣閩南語口語語料庫. 國立中正大學語言學研究所
- Radio broadcast shows. 28 hours. 315,069 tokens.

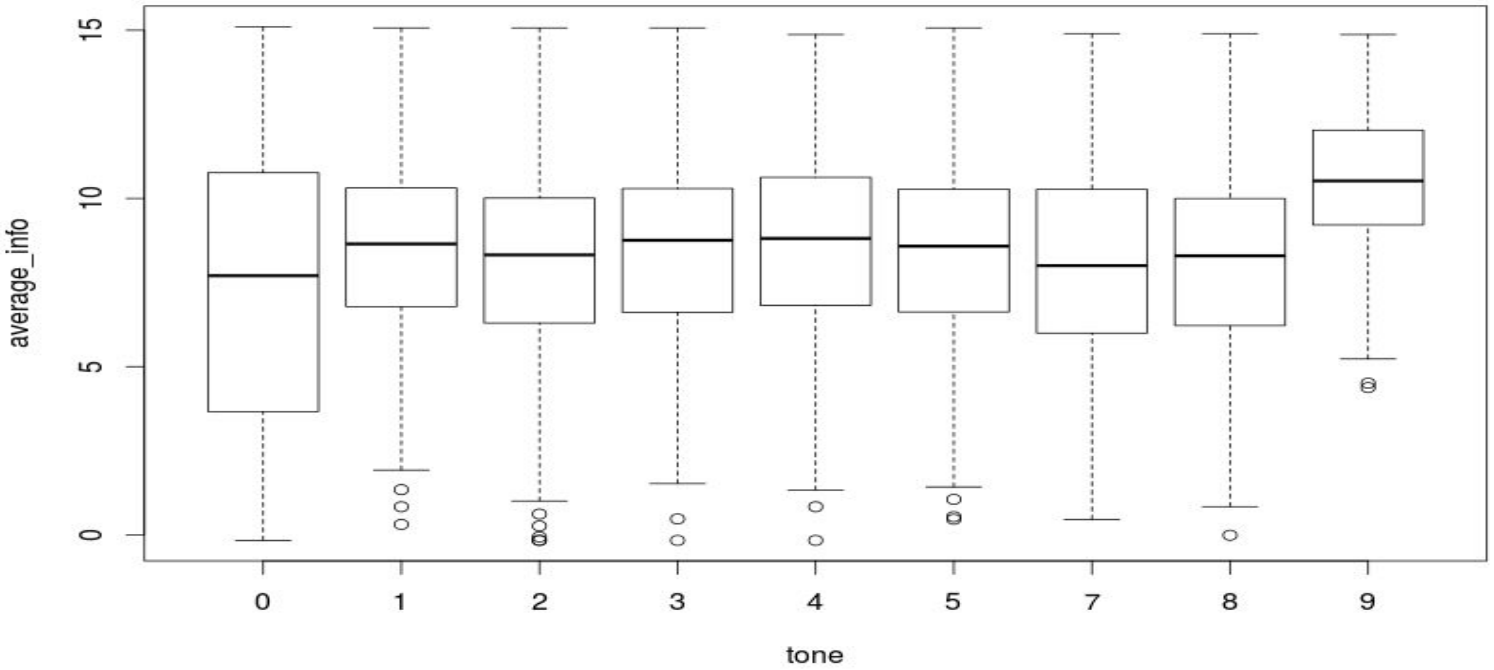
Participants: 001(host), 002(hostess),
Coder: Pan
Filename: RB001
Language: Taiwanese
Date: 4-Oct-1999
Tape Location: RB D1, Tracks 1-5
Time Duration: 32'
Activities/Comments: Radio show
Track 1-1

002: 親愛<chin1ai3> e0<e0> 聽眾<thiann1ciong3> 朋友<kong2li2> 逐家<tak8ke1> 好<ho2>。
001: hai0<hai0>, 逐家<tak8ke1> 好<ho2>。
002: henn0<henn0>。
002: 歡迎<huan1ging5> 再度<cai3too7> 收聽<siu1thiann1> 咱<lan2> e0<e0> 節目<ciat4bok8>。
001: 是<si7>, [遮2<cia1> 是<si7> i] ,
002: [我<gua2> 是<si7> i] ,
001: henn0<henn0>, 雲嘉<hun5ka1> 廣告<kong2ko3> 電台<tian7tai5> hoonn0<hoonn0>,
002: henn0<henn0>。
001: 咱<lan2> 來<lai5> 進行<cin3hing5> 每<mui2> 禮拜一<le2pai3it4> 到<kau3> 拜六<pai3lak8>, 五點<goo7tiam2> 至<ci3> 六點<lak8tiam2> e02<e0>。
002: henn0<henn0>。

Hokkien Tonal System

Name	Tone Value	Example
Tone1	44	taŋ1 ‘east 東’
Tone2	53	taŋ2 'executive 董’
Tone3	21	taŋ3 'frozen 凍’’
Tone4	32 (checked)	tak4 ‘touch 觸’
Tone5	24	taŋ5 ‘copper 銅’
Tone7	22	taŋ7 ‘move 動’
Tone8	4 (checked)	tak8 ‘subsequently 逐’
Tone0 (neutral tone)	1 or context dependent	khih0 '(get) in (進)去' See Myers & Li (2009)
Tone9 (Contraction)		[siang9] ‘who 啥人’ derived from [siánn-lâng] See 歐& 蕭 (1997)

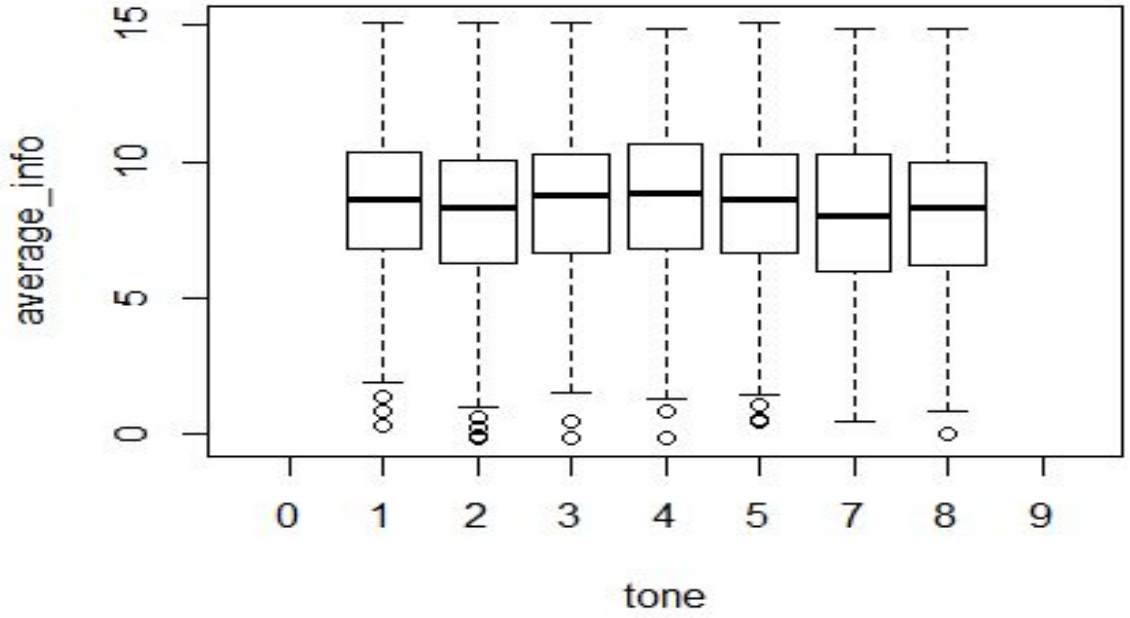
Result Hokkien *average info: ANOVA (F(8,2225) =6.961, p <.0001)*



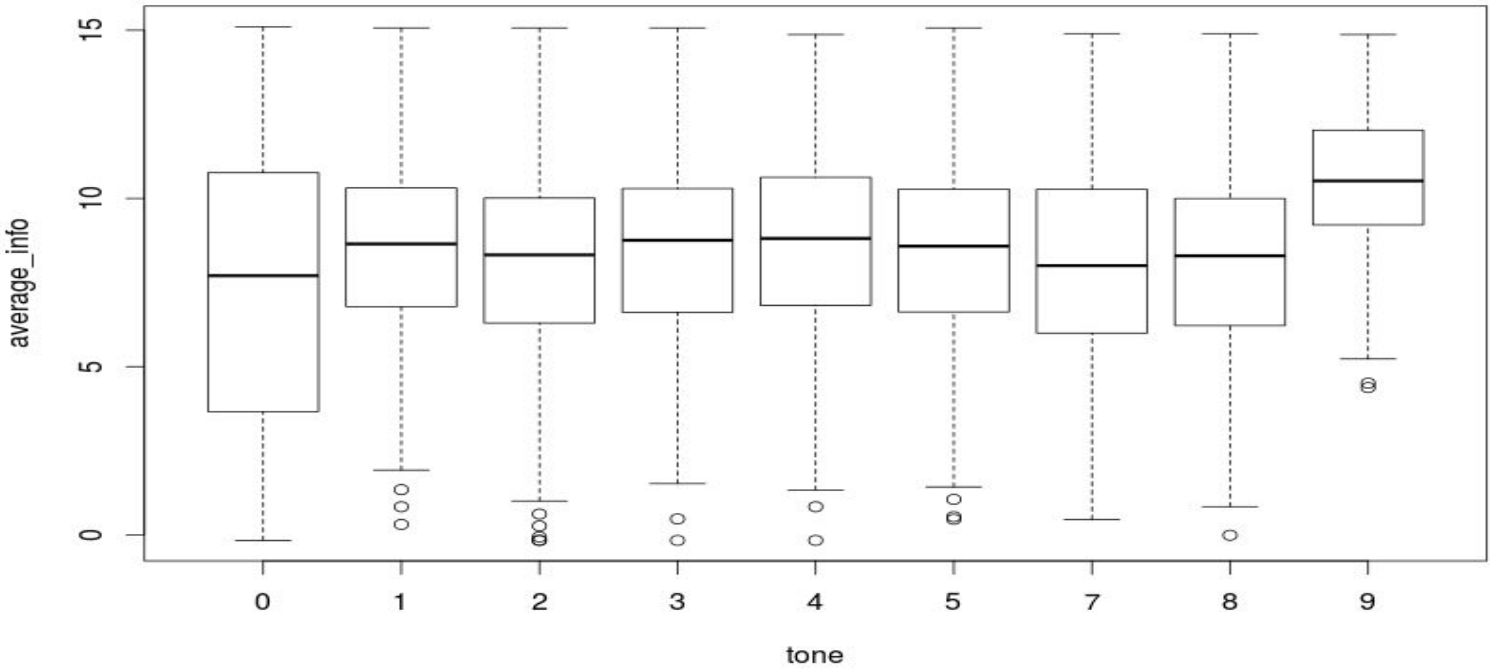
Tukey hsd post hoc test: significant only when 0 or 9 is compared

	Diff	lwr	upr	p
9-0	3.118172007	1.66428956	4.5720545	0
9-7	2.302086937	0.9232013	3.6809726	0.0000084
9-2	2.277937062	0.91010089	3.6457732	0.000009
9-8	2.223165159	0.76830069	3.6780296	0.0000776
4-0	1.378082338	0.44243173	2.3137329	0.0001752
9-1	1.9450933	0.58286125	3.3073254	0.0003327
9-5	1.944666164	0.57190814	3.3174242	0.0003894
9-3	1.937758987	0.55550852	3.3200095	0.0004774
1-0	1.173078707	0.32862887	2.0175285	0.0005683
5-0	1.173505843	0.31217898	2.0348327	0.0008152
3-0	1.18041302	0.30403655	2.0567895	0.0010017
9-4	1.740089669	0.31951841	3.1606609	0.0046306
2-0	0.840234945	-0.01322575	1.6936956	0.0575755
7-0	0.81608507	-0.0549746	1.6871447	0.0870801

Result Hokkien (w/o t0 and t9) *average info: ANOVA (F(6,2005) =1.589, p =0.146)*



Result Hokkien (average info): 0 < {1~8} < 9



Discussion

- Tone9 (contraction) is a combination of two morphemes. Thus it should have more information content.
- Tone0 (neutral tone) is short and light phonetically. Thus it should less information content.

- Why do the other tones (T1 toT8) have the same information content, while in Mandarin different tones have different amount of information content (i.e. T1>T3>T2>T4)?

Contact:
Yuan-Lu Chen
Univ. Arizona
cheny@email.arizona.edu

References

- Chen, M. Y. 2000. Tone Sandhi: Patterns across Chinese Dialects. Cambridge: Cambridge University Press.
- Duanmu, S. 2000. The Phonology of Standard Chinese. Oxford: Oxford University Press.
- Glenn, M., & Linguistic Data Consortium. (2013). GALE phase 2 Chinese broadcast news transcripts. Philadelphia, PA: Linguistic Data Consortium.
- Li, Charles N. and Thompson, Sandra A. (1977). The acquisition of tone in Mandarin-speaking children. Journal of Child Language, 4, pp 185-199. doi:10.1017/S0305000900001598.
- Moore, C.B., and Jongman, A. (1997). Speaker normalization in the perception of Mandarin Chinese tones. Journal of the Acoustical Society of America 102, 1864-1877.
- Myers, J., & Li, Y. (2009). Lexical frequency effects in taiwan southern min syllable contraction. Journal of Phonetics,37(2), 212-230. doi:10.1016/j.wocn.2009.02.002
- Piantadosi, S., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. Proceedings of the National Academy of Sciences, 108(9), 3526.
- Pypinyin. (2014). Retrieved May 3, 2015, from http://pypinyin.readthedocs.org/en/latest/index.html#api
- 歐淑珍, & 蕭宇超. (1997). 從 [韻律音韻學] 看台灣閩南語的輕聲現象. 聲韻論叢, (6), 865-895.