An estimation of functional load from the frequency of phonological oppositions

Mafuyu Kitahara
(Waseda U)

Phonology Forum 2005 @ Fukuoka U
Aug. 25

Introduction

- Language is a system of differences (de Saussure 1910).
- Differences implemented in minimal pairs.
- How many minimal pairs possible?

$$C = \sum_{i=1}^{m} \frac{mn^m(n-1)}{2}$$

m: max length
n: inventory size

Explosion

- For example...
  - *n*=5, *m*=1 $$C = \frac{mn^m(n-1)}{2} = \frac{1 \times 5^1(5-1)}{2} = 10$$
  - *n*=5, *m*=2 $$C = 10 + \frac{2 \times 5^2(5-1)}{2} = 110$$
  - *n*=5, *m*=3 $$C = 10 + 100 + \frac{3 \times 5^3(5-1)}{2} = 860$$
- A little more realistic *n* & *m* ... BANG!
  - *n*=20, *m*=10 ... *C* = 2,037,221,052,631,580
    (2 quadrillion 37 trillion...)

Lexicon as suppressor

- No natural language uses *n* phonemes evenly nor maximally.
- Minimal pairs are only viable in a lexicon.
- Lexicon SUPRESSES the potential of the system of differences.
- Phonology is all about why unevenness and demaximization come in.
  e.g. phonotactics, alternation...source of unevenness

Functional load

- How uneven? .... not a new question.
- Classic functional load (Martinet 1962; Hockett 1967):

$$F(/\text{a}/,/\text{b}/) = F(L^m) - F(L^{m-1})$$

$F(/\text{a}/,/\text{b}/)$: functional load of /a/,/b/

$F(L)$: functional load of the system L

L^m: System of m phonemes

L^{m-1}: System of $m-1$ phonemes

- *F(L)*=entropy? No simulation provided.

Quantification of FL

- Surendran & Niyogi (2003):
  - Reinterpret Hockett's formulation.
  - Probability estimation of *n*-gram in a corpus.
  - Can compare segmental as well as prosodic units (Surendran & Levow 2003).
- Problem:
  - Depends heavily on corpus.
  - *n*=1 in *n*-gram make little difference = just frequency effects?

Method (1)

- Count the number of minimal pairs.
  - Neighborhood calculation in NTT database: 80,000 words (Amano & Kondo 1999).
  - Original neighborhood (Greenberg & Jenkins 1964):
    - Substition = pit - bit
    - Addition = pit - spit
    - Deletion = pit - it
  - Keep the number of morae for Japanese data
    - Licit addition = tango – tanago, aato - paato
    - Illicit addition = paato - apaato

Method (2)

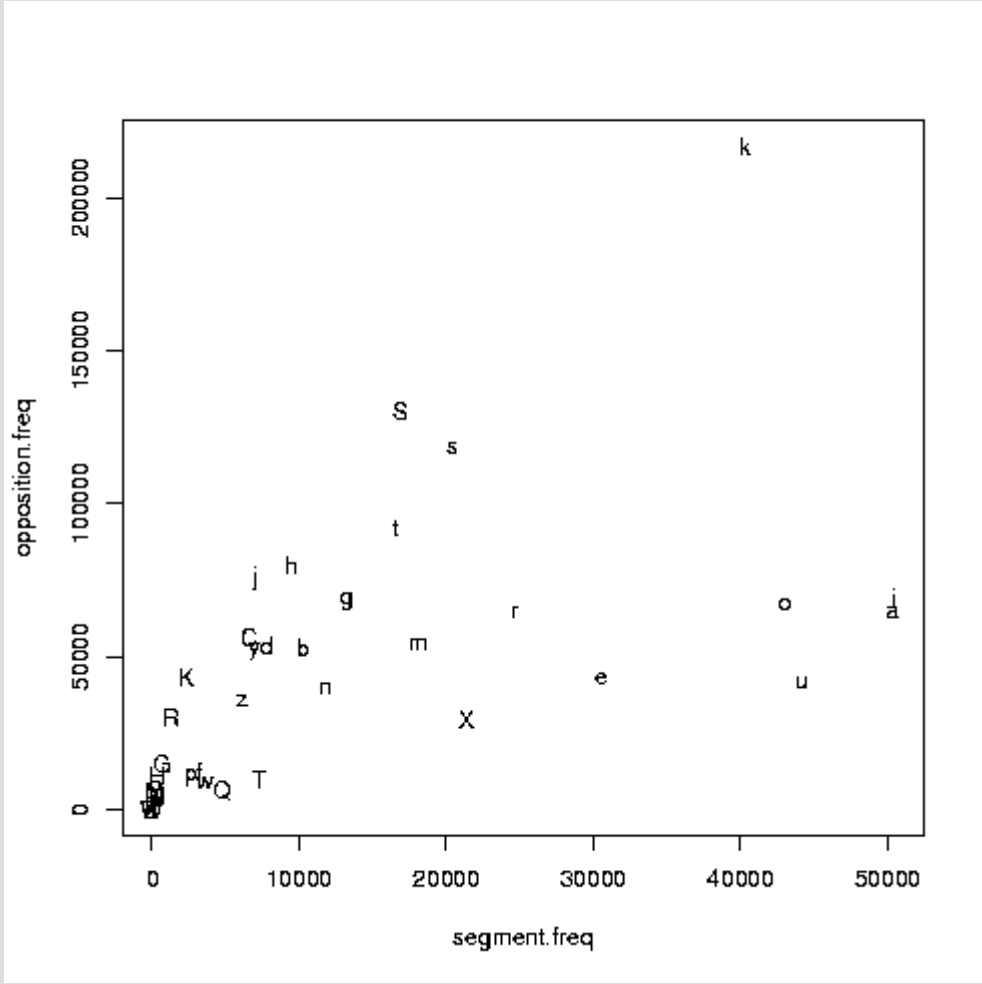
- Count the number of oppositions for each segment and each feature.
  - Feature system based on Halle (1992).
  - Oppositions involving unspecified nodes or features are not counted.
- All done by perl scripts (available from the author if interested).

Data (1)

- Top 10 list of opposition freq and segment freq.

Opposition		Segment	
216650	k	50414	i
130418	ʃ	50279	a
118410	s	44077	o
92041	t	43099	u
79662	h	40285	k
75523	j	30508	e
69138	i	24698	r
68520	g	21374	N
67109	o	20411	s
64552	r	18075	m

- Opposition freq against segment freq.



Discussion (1)

- /k/ is by far the most useful segment in Japanese.
- Vowels appear frequently but not very helpful for distinction.

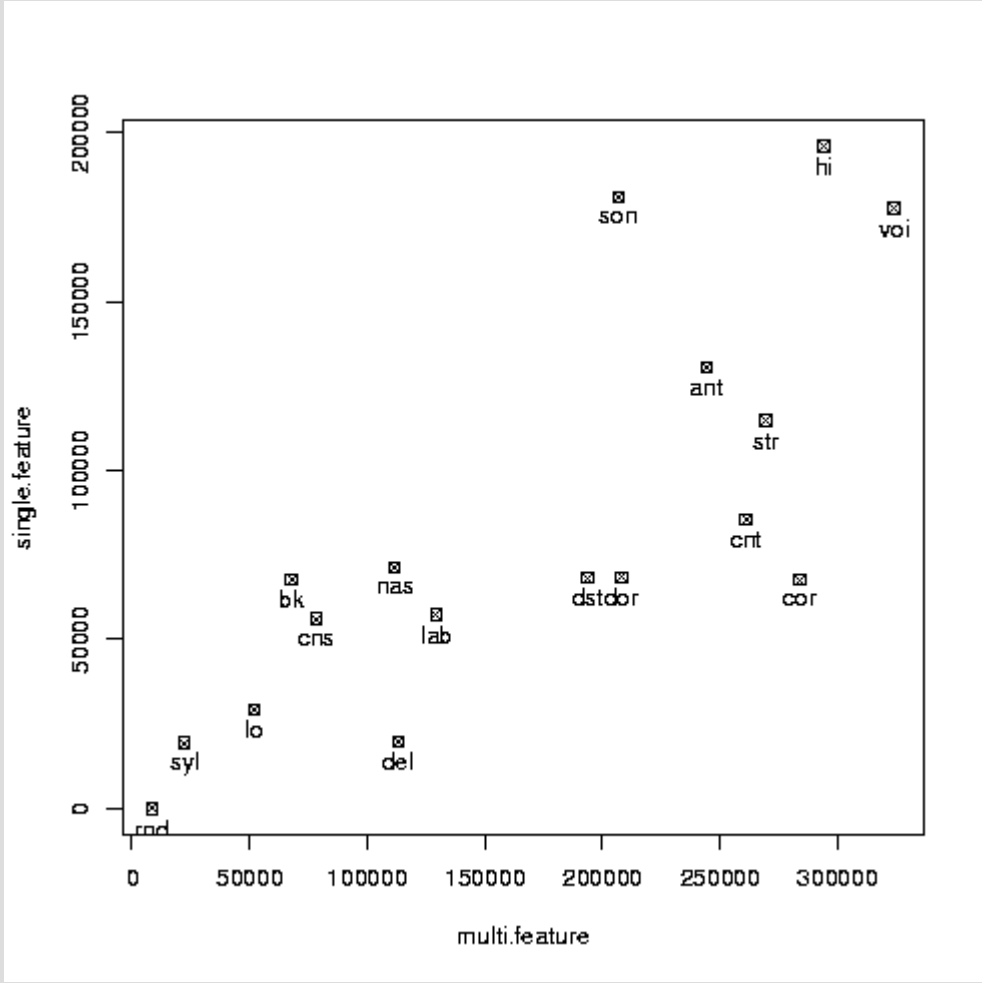
- Word(token) frequency and/or familiarity can be incorporated (in the future).

Data (2)

- List of single feature and multi-feature opposition.

	Single		Multi
hi	196016	voi	323943
son	180988	hi	294106
voi	177694	cor	283914
ant	130565	str	269274
str	114892	cnt	261072
cnt	85537	ant	244144
nas	71298	dor	208015
dor	68491	son	206665
dst	68301	dst	193698
bk	67756	lab	129349
cor	67756	del	113022
lab	57451	nas	111799
cns	56106	cns	78317
lo	29266	bk	67756
del	19742	lo	51794
syl	19402	syl	22211
rnd	0	rnd	8446

- Single feature vs multi-feature opposition.



Discussion (2)

- Single feature opposition

u	+	−	+	0	0	0	+	−	−	−	0	0	0	0	+	−	+
i	+	−	+	0	0	0	+	−	−	−	0	0	0	0	+	−	−
	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=	1

- Multi-feature opposition

p	−	+	−	−	−	−	−	−	+	−	−	−	0	−	−	−	−
d	−	+	−	−	−	−	+	−	−	−	+	+	−	−	−	−	−
	=	=	=	=	=	=	1	=	1	=	1	1	0	=	=	=	=

- Height and voicing are the two major sources of distinction.

Summary and implication

- Lexicon as suppressor of the system of differences.
- Frequency of oppositions, segments, and features provide a nice summary of HOW lexicon distorts the system.
- Yet another clue for universal markedness hierarchy.

Thank you