ON THE INTERPRETATION OF CONSONANT-VOWEL CO-OCCURRENCE FREQUENCY BIASES

Eleonora C. Albano

State University of Campinas, Campinas, São Paulo, Brazil
albano@unicamp.br

ABSTRACT

Certain recurrent consonant-vowel co-occurrence frequency biases have often been attributed to the biomechanics of the mandible. This paper takes issue with this claim by pointing to some methodological flaws in the literature supporting it.

CV co-occurrence data on Spanish and Portuguese are brought to bear on such neglected issues as sampling, effect size, and count type. It is shown that small samples are unrepresentative, effect size is generally low, and type and token counts lead to different results.

These drawbacks notwithstanding, the paper supports the need for further research into the biomechanical bases of such biases. Two linguistic contexts have been uncovered where they prove to be statistically robust in both languages: initial unstressed position; and the lexical subset of words with repeated CV pairs. Both contexts have been related to phonetic complexity in the literature.

Thus, bias motivation may lie not simply in biomechanics, but, rather, in its interaction with language specific, linguistic proper constraints.

1. INTRODUCTION

Jansen [2] opened the CV co-occurrence literature by reporting on five languages which consistently favor certain CV pairs. Challenging earlier work on other segment classes [6], he attributed such biases to articulatory economy. Later, Maddieson & Precoda [5] contributed more data and spelled out some other possible influences on the biases.

Meanwhile MacNeilage & Davis [3] began to investigate the same problem in child language. They found that the following CV pairs are favored in both babbling and first words of children acquiring a number of languages: labial C/central V, coronal C/front V, dorsal C/back V. This is the basis of their Frame-then-Content Theory (henceforth FC), which claims that the biases in question originate in mandible oscillations framing articulator motion so that C and V targets are hit at first unintentionally in passing.

MacNeilage & Davis's ideas about the possible role of biomechanics in the acquisition of phonology are important and innovative, but should not be confused with their reductionist stance on the biomechanics of the syllable [4].

There is no a priori reason why the biases in question, if indeed true, should originate in the mandible, as opposed to the rest of the vocal tract. In addition, the authors offer no explicit account of how the early indivisible stomato-gnathic mandible oscillations should develop into C and V gestures, generally assumed to be independent [1, 2, 5].

This paper addresses this issue by returning to the original linguistic perspective, focused mainly on languages and lexicons [2, 5, 6]. The following questions are pursued:

- Are such biases found in large lexical samples?
- If so, what is the size of the statistical effect?
- If stronger effects do occur, where do they appear?

Statistics absent from the CV co-occurrence literature in spite of being standard in other areas [7] are brought to bear on these issues. Their use strongly supports the biases only in certain linguistic contexts on which other interactions with biomechanics have been reported [8].

2. METHODOLOGY

Choice of Spanish and Portuguese was solely determined by data availability, since the questions under discussion could be studied in any language.

A familiar bias size index, the O/E ratio, is avoided here because of its inaccuracy. Though chi square remains indispensible for treating nominal data [7], its low statistical power requires stricter cell significance criteria in order to minimize Type I errors.

2.1. Corpora

Due to technological limitations, previous studies have used small databases, varying from 2,000 to 10,000 words. Two easily accessible large corpora secure representativeness here:

2.1.1. Portuguese Lael-Fala (henceforth, LAEL)

This is a free, orthographically transcribed, 45,000 word corpus, containing lectures, interviews, and conversations of educated São Paulo residents, available at: http: //www2.lael.pucsp.br/corpora/.

2.1.2. Spanish Call Home (henceforth, CHome)

This is a phonetically transcribed, 45,000 word corpus containing telephone conversations of Spanish speaking US residents, available on demand from *Linguistic Data Consortium*.

2.2. Corpus treatment

After removal of foreign words and acronyms, orthography-to-phone conversion was performed on LAEL with software from our own laboratory. LDC's transcription of CHome was supplemented with stress marks.

2.3. Sampling

Both corpora consist of word lists with frequencies. Type and token frequencies can thus be easily calculated for any given CV pair.

Smaller, random and nonrandom 2,000 word samples were drawn from the types of both corpora to assess the effects of corpus size. The nonrandom samples were drawn sequentially starting at a randomly selected point in the list.

2.4. Coding

All CV pairs were coded as to the class of their respective consonants and vowels, namely: labial, coronal, and dorsal; and front, central, and back. They were also coded as initial or medial, and stressed or unstressed.

2.5. Frequency Counts

For the full corpora, frequency counts of the cross-tabulated C and V categories were nested into the layers 'position' and 'stress'. Nesting was not computed for small samples.

Each corpus went through an additional count on the word subset containing any repeated (not necessarily reduplicated) CV pair.

2.6. Statistics

Two statistics supplement chi square: a measure of association strength, namely, Cramer's V; and a measure of cell significance, namely, Sokal & Rohlf's critical values for Freeman-Tukey deviates [7]. Significance level is set by the Bonferroni criterion for multiple comparisons. Where applicable, factors are probed with a log linear analysis of the contingency table.

3. RESULTS

The support provided by the data to the universality of the observed co-occurrence patterns is weak overall, but considerably stronger locally.

3.1. Types and tokens

Work on FC generally compares lexicons with child or adult corpora [4], ignoring that the type/token distinction is implied in the counts.

Tables 1-3 show that the two kinds of counts yield very different results. Although the X^2 and p values in Table 1 attest the existence of biases in both cases, the V values indicate, in turn, that the association strengths are quite low[1] in types, and negligible in tokens. This calls into question the validity of equating types to tokens.

Table 1: X^2, p, and Cramer's V for types and tokens in Call Home and LAEL.

	Types (N≈130,000)			**Tokens (N>3,000,000)**		
	X^2	p	V	X^2	p	V
CHome	5627	0.000	0.15	311087	0.000	0.04
LAEL	2846	0.000	0.10	36903	0.000	0.07

This impression is reinforced by Tables 2 and 3, where the significant biases for types and tokens are cross-tabulated for each corpus. Note that the biases disperse among 6/9 cells. FC predictions are henceforth highlighted in grey.

Table 2: Significant biases for types and tokens in Call Home.

C Home	**Front**	**Central**	**Back**
Labial	Types	Tokens	Tokens
Coronal	Types Tokens		
Dorsal		Types Tokens	Types

Table 3: Significant biases for types and tokens in LAEL.

LAEL	**Front**	**Central**	**Back**
Labial	Types	Types Tokens	
Coronal	Types Tokens	Types	
Dorsal		Tokens	Types Tokens

[1] Usually, V<0.2 is deemed low; and V<0.1, negligible.

The coincidence ratio among type and token biases is 4/8 in CHome, and 6/9 in LAEL. Biases are, therefore, inconsistent across the two counts, rendering comparison innocuous.

Incidentally, FC predictions display exactly the same coincidence ratios (namely, 4/8 and 6/9, respectively). Note that, overall, LAEL does slightly better than CHome, partly agreeing with the predictions in both types and tokens.

In any case, even if LAEL weakly supports FC, an effect size of less than 20% (i.e., V < 0.2) is not very encouraging. Besides, the fact that the present V values rest on very large samples suggests that the associations involved may indeed be inherently weak.

Finally, the low coincidence ratio between the two languages should not go unnoticed: it is just 3/9 for types, and 3/8 for tokens. This says that even closely related languages need not share exactly the same biases. The fit with FC is equally inconclusive: it is attained in 10/17 cases.

3.2. Sample size

Trying to study CV co-occurrence frequencies with small samples may in fact be misleading. Tables 4-6 show why.

Table 4: X^2, p, and Cramer's V for random and nonrandom samples of Call Home and LAEL.

	Random (N=2,000)			**Nonrandom (N=2,000)**		
	X^2	p	V	X^2	P	V
CHome	224	0.000	0.14	231	0.000	0.15
LAEL	155	0.000	0.12	163	0.000	0.12

Table 5: Significant biases for random (R) and nonrandom (NR) samples of Call Home.

C Home	**Front**	**Central**	**Back**
Labial			NR
Coronal	R/NR	NR	
Dorsal		R/NR	R

Table 6: Significant biases for random (R) and nonrandom (NR) samples of LAEL.

LAEL	**Front**	**Central**	**Back**
Labial	NR		
Coronal	R		NR
Dorsal		R/NR	R

The inconsistency and dispersion of the biases in the random and nonrandom 2,000 word samples are very clear above. In addition, V values are not as low as might be expected, suggesting that the criterion for sufficient association strength in this kind of study should be 0.2, not 0.1. Note that, under such a criterion, even the large sample biases should be interpreted with caution.

3.3. Linguistic context

An ensuing question is: are there lexical settings where CV co-occurrence biases become stronger and more consistent? The answer is positive and points to their language dependent facet.

3.3.1. Position and Stress

Inasmuch as their influence ranges from segment inventories to phonetic detail, position and stress are possible linguistic effects on CV phonotactics. As shown in Tables 7-8, nesting the co-occurring C and V pairs into stress and position layers brings out a stronger effect which remains consistent across languages. Note the V values in **bold** italics.

Table 7: X^2, p, and Cramer's V for the position and stress lexical subsets of Call Home.

	CHome	**X^2**	**p**	**V**
	Str.	150	0.000	0.15
Initial	***Unstr.***	*6512*	*0.000*	***0.37***
	Str.	552	0.000	0.09
Medial	**Unstr.**	1791	0.000	0.11

Table 8: X^2, p, and Cramer's V for the position and stress lexical subsets of LAEL.

	LAEL	**X^2**	**p**	**V**
	Str.	260	0.000	0.16
Initial	***Unstr.***	*5193*	*0.000*	***0.33***
	Str.	823	0.000	0.11
Medial	**Unstr.**	1560	0.000	0.11

Now note the italicized cells in Tables 9-10.

Table 9: Significant biases for the position and stress lexical subsets of Call Home (significance = S).

				V class		
Position	**Stress**	**Call Home**		**F**	**C**	**B**
Initial	**Str.**	**C class**	**L**		S	
			C	S		
			D			S
	Unstr.	***C class***	***L***		***S***	
			C	***S***		
			D		***S***	***S***
Medial	**Str.**	**C class**	**L**		S	S
			C			S
			D		S	
	Unstr.	**C class**	**L**	S		
			C	S		
			D			S

Table 10: Significant biases for the position and stress lexical subsets of LAEL (significance = S).

Position	Stress	LAEL		V class		
				F	**C**	**B**
Initial	**Str.**	**C class**	**L**			
			C	S		
			D		S	S
	Unstr.	***C class***	***L***		***S***	
			C	***S***		
			D			***S***
Medial	**Str.**	**C class**	**L**	S		
			C			
			D		S	S
	Unstr.	**C class**	**L**	S		
			C	S		
			D			S

Summing up, Tables 7-8 finally show strong enough, though moderate, effects in initial unstressed position; while Tables 9-10 show that 6/7 biases therein cohere with FC predictions.

For each corpus, a log linear analysis was run to probe into factor interaction[1]. In both cases, all factors and interactions turned up significant. Position and stress are thus intertwined with phonetic content in determining CV co-occurrence.

This finally leads to the following question: what attracts the favored biases to initial unstressed position? Let us leave this open until we look into another setting where the same biases show up with sufficient association strength.

3.3.2. Repetition of CV Pair

Since 'initial' and 'unstressed' are contradictory positions with respect to strengthening, it may be instructive to look at other similarly complex environments. Another documented source of complexity is repetition of segment sequences [8]. Thus, the lexical subset consisting of all words containing repeated CV pairs may be a good bet.

Here, again, V values are above 0.2 and biases agree almost perfectly with FC predictions. By the way, the recurrence of the diverging dorsal/central bias (cf. Tables 2-3 and 9) is noteworthy.

Table 11: X^2, p, and Cramer's V for the lexical subsets of Call Home and LAEL with repeated CV's.

	Repeated CV (N ≈ 1,500)		
	X^2	p	V
CHome	234	0.000	**0.27**
LAEL	227	0.000	**0.28**

[1] Results are not shown here for lack of space.

Table 12: Significant biases for the lexical subsets of Call Home and LAEL with repeated CV's.

C Home	**Front**	**Central**	**Back**
Labial		CHome LAEL	
Coronal	CHome LAEL		
Dorsal		CHome LAEL	CHome LAEL

The most reasonable interpretation for these facts seems to be that the biases recur in an attempt to reduce complexity. This implies that the favored CV pairs are the least complex among the nine possible combinations. Why should this be?

4. CONCLUSIONS

FC supporters generally attribute bias simplicity to "frame dominance". But it is not at all clear how such biases could arise solely from mandible oscillation. The only warranted inference here is that they may be a means of reducing complexity.

Inasmuch as stress, position and repetition are related to articulatory effort, an attempt to pull these findings together would read approximately as follows: lexicons incorporate means of reducing complexity into CV phonotactics. To pursue the biomechanics behind such a trend, we must inquire whether the effect uncovered here recurs in other languages, and, if so, seek explanation in explicit, computational models of gestural coordination [1].

5. REFERENCES

[1] Browman, C., Goldstein, L. 1992. Articulatory phonology: an overview. *Phonetica*, 49, 155-180.

[2] Jansen, T. 1986. Cross-linguistic trends in the frequency of CV sequences. *Phonology Yearbook*, 3, 179-195.

[3] MacNeilage, P. F., Davis, B. L.. 1990. Acquisition of speech: frames, then content. In: M. Jeannerod (ed.) *Attention and Performance, vol. XIII*. Hillsdale, NJ: Erlbaum, 452–468.

[4] MacNeilage, P. F., Davis, B. L. 2000. On the origin of internal structure of word forms. *Science*, 288, 527-531.

[5] Maddieson, I., Precoda, K.1992. Syllable structure and phonetic models. *Phonology*, 9, 45-60.

[6] Ohala, J. J., Kawasaki, H. 1984. Prosodic phonology and phonetics. *Phonology Yearbook*, 1, 113-127.

[7] Sokal, R. R., Rohlf, F. J. 1995. *Biometry*. San Francisco, CA: W. H. Freeman and Company.

[8] Walter, M. A. 2007. *Repetition avoidance in human language*. MIT doctoral dissertation.