Análise quantitativa da freqüência dos fonemas e estruturas silábicas portuguesas

Mário Eduardo Viaro[1], Zwinglio O. Guimarães-Filho[2]

[1]Departamento de Letras Clássicas e Vernáculas, FFLCH (USP)

[2]Instituto de Física (USP)
maeviaro@usp.br, zwinglio@usp.br

Abstract. *Statistical proprieties of the structure of the Portuguese syllables and the combination of the phonemes are considered. The corpus is based on a mechanical transcription of a large list of Portuguese words. In the CV structure, some clear preferences and rejections between C and V are perceptible by means of the rank analysis.*

Key-words. *Phonology; Portuguese language; frequency; syllables.*

Resumo. *Discutem-se características estatísticas da estrutura das sílabas portuguesas e da combinação de fonemas. O corpus é baseado numa transcrição mecânica de uma grande lista de palavras portuguesas. Na estrutura CV, algumas preferências e rejeições evidentes entre C e V são perceptíveis por meio da análise de posto.*

Palavras-chave. *Fonologia; língua portuguesa; freqüência; sílabas.*

0. Introdução

Nos textos teóricos, principalmente nos de caráter dedutivo, fazem-se, com freqüência, algumas generalizações que se tornam premissas de ampla argumentação. A verificação quantitativa dos fenômenos em que se fundam essas generalizações é, sem dúvida, bastante necessária. Tal tarefa é, ocasionalmente, atribuída aos que testam a validade científica. Em lingüística, esse teste é comumente feito com *corpora,* dos quais se servem os pesquisadores, quer para corroborar o que se afirma, quer para romper o que se pensava erroneamente de maneira intuitiva. Todo *corpus,* no entanto, tem suas limitações, mas é verdade que quanto mais extenso ele for, mais consideráveis são os dados que podem fornecer.

O nosso *corpus* compõe-se de 150.875 palavras, todas existentes no Dicionário Houaiss da Língua Portuguesa. Trata-se, portanto, de um *corpus* da língua escrita, sem consideração à freqüência de uso das palavras e composto de verbetes não-lematizados. Foram deixadas de lado as abreviações, siglas, elementos de composição, estrangeirismos evidentes, bem como os homônimos e as palavras hifenizadas, formadas por justaposição. Cumpre observar que os resultados extraídos desses dados se baseiam em um *corpus* bem maior que o apresentado em trabalhos parecidos (cf. SILVA *et alii*, 1993)

Sobre essa lista de palavras, trabalhou-se com programas computacionais feitos na plataforma *MatLab* (*The MathWorks, Inc.,* 1984-2002). Inicialmente, separaram-se

as sílabas, em seguida, localizou-se a tônica, transcreveu-se num alfabeto fonológico intermediário (com 65 caracteres distintos, representando fonemas e arquifonemas, especificados por seu contexto fonotático), construído para abarcar uma grande gama de variações fonéticas da língua portuguesa e, por fim, fez-se uma simulação de interpretação fonética para algumas variantes do português. A apresentada aqui seria uma interpretação da transcrição fonológica sob uma variante "paulista", ou seja, que tem [s], [w] e [ɾ] nas codas, assim como [tʃ] e [dʒ] antes de [i]. Para casos complexos de transliteração, como os oferecidos pelos grafemas *x, e, o,* utilizou-se a ortoépia fornecida pelo dicionário.

Antes da interpretação fonética, a transcrição fonológica (ou base transcrita, doravante BT) podia ter as opções de prever casos de adição de fonemas (prótese, epêntese e paragoge, doravante BE), assim como fusão de sílabas por sinérese (BS). Adições e sinéreses juntas também podiam formar uma outra base, doravante BES.

Todos os resultados que se seguem, portanto, são associados às limitações que a metodologia impôs. As separações e transcrições fonéticas foram submetidas, por amostragem, a uma verificação manual e, por meio dela, pode-se estipular que a margem de erro é inferior a 0,5%.

1. Tamanho das palavras

As palavras do *corpus* possuem de 1 a 45 fonemas (ou 46 na BE). Isso equivale a dizer que têm de 1 a 20 sílabas (ou 21 na BE). A maior palavra é *pneumoultramicroscopicossilicovulcanoconiótico.* A quantidade de palavras por fonema cresce até 9 fonemas (que registra cerca de vinte e três mil verbetes) e em seguida cai. Com relação às sílabas, os tetrassílabos formam o maior grupo de palavras (por volta de quarenta e sete mil verbetes), como se vê pela tabela abaixo:

Tabela 1. Distribuição de palavras portuguesas segundo a quantidade de sílabas

Núm. Sílabas	BT	BE	BS	BES
1	546	467	790	704
2	11712	11274	13165	12698
3	36790	35742	38113	36957
4	48218	47127	47205	46174
5	33125	33231	31982	32148
6	13926	15116	13355	14540
7	4665	5483	4503	5345
8	1440	1807	1354	1733
9	362	489	324	447
10	76	97	71	91
>10	16	43	14	39

Cumpre observar que, diante da possibilidade combinatória dos fonemas, os números acima são sempre pequenos. Por exemplo, os monossílabos correspondem a um mínimo de 467 formas (BE) e a um máximo de 790 formas (BS). Respeitando todas as limitações combinatórias entre ataque, núcleo e coda, o português seria capaz de produzir 4627 seqüências monossilábicas (VIARO 2005). As palavras de monossílabos

reais, no *corpus,* portanto, equivalem a um percentual relativamente reduzido (respectivamente 10,1% e 17,1%). Um estudo de freqüência de uso reduziria ainda mais esse percentual.

2. Tonicidade

É de especial importância para os estudos de acento lexical partir da hipótese de que o português seja uma língua de acentuação paroxítona. Nossos dados confirmam isso. No entanto, a proparoxítona, caso tido como excepcional no português, muitas vezes não tem porcentagens tão desprezíveis como se pensa.

Tabela 2. Distribuição das palavras portuguesas conforme a posição da sílaba tônica

Sílaba tônica	BT	BE	BS	BES
última	37591	37496	37602	37514
penúltima	94326	93604	94164	93443
antepenúltima	18413	18652	18320	18558

As oxítonas equivalem a 25%, as paroxítonas a cerca de 62% e as proparoxítonas a 12% do total. Não estão nesta listagem os monossílabos (0,3% a 0,5%). O interessante é que palavras como *rítmico* podem, na BE e na BES, ter acento tônico na pré-antepenúltima sílaba. Este caso, longe de ser excepcional, ocorre em 653 palavras (sobretudo terminadas em *-ptero, -óxido* etc. e 3,3% das terminadas em *–ico* com *i* átono). Isso equivale a 0,4% da BE, uma quantidade maior do que a de monossílabos. Há quatro palavras nas mesmas bases que têm acento tônico na pré-pré-antepenúltima sílaba: *arqueópterix, diatenópterix, dípterix, monópterix.* Com base em buscas realizadas no Google em 15/8/2006 restritas a páginas em português pudemos constatar que estas palavras têm freqüências de uso extremamente raras: apenas 20 ocorrências de páginas contendo a palavra *arqueópterix*, e 1 ocorrência para *dípterix* foram encontradas na base de cerca de 500 milhões de páginas em português do Google.

Outro resultado interessante é que quanto mais longa a palavra, maior a tendência de ela ser proparoxítona do que oxítona, conforme se pode observar na Figura 1.

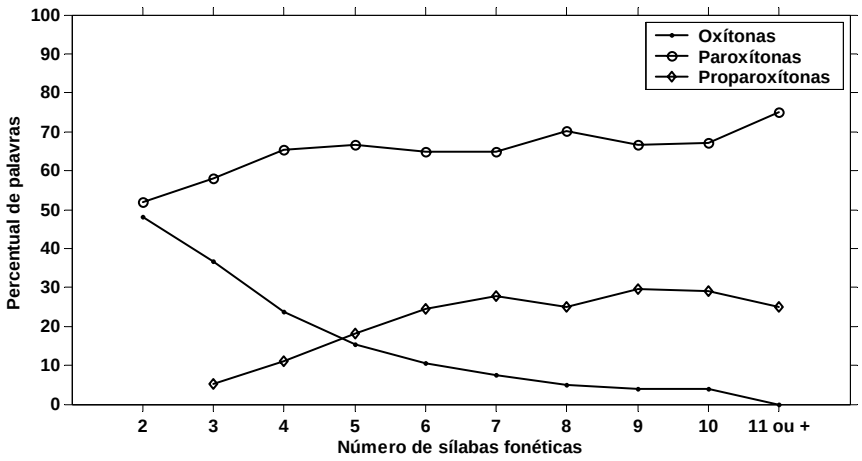

Figura 1 - Relação entre tonicidade e número de sílabas fonéticas.

Segundo a Figura 1, paroxítonas e oxítonas têm basicamente o mesmo percentual nos dissílabos, ao passo que o número de oxítonas pentassílabas é menor do que o de proparoxítonas, mantendo-se abaixo delas, sempre em queda, à medida que o número de sílabas aumenta.

Os dados permitem ainda questionar alguns modelos, como o da consoante abstrata nas oxítonas finalizadas em vogal e o da extrametricidade (BISOL 1992). Apesar de a maioria das oxítonas terminarem em consoante e das paroxítonas, em vogal, 5% das oxítonas (BT) terminam em alguma vogal não-nasal e 12% das paroxítonas terminam em consoante, glide ou em vogal nasal, isso sem falar do nada desprezível número de proparoxítonas (0,5% termina em consoante e 16% em vogal).

Tabela 3. Distribuição de palavras portuguesas conforme a classe do último fonema

	-C	-V
Monossílabos	370	176
Oxítonas	32255	5336
Paroxítonas	4431	89894
Proparoxítonas	182	18231
Total	37238	113637

Se analisarmos mais de perto os dados da tabela 3, verificaremos que se a consoante em questão for um /N/, as distâncias entre as paroxítonas e as oxítonas diminuem: 51,6% dessas palavras são oxítonas, mas 43% são paroxítonas. Com /S/ a distância é ainda menor: 45% são indiferentemente oxítonas ou paroxítonas. Com /L/, 61% são oxítonas, mas 39% são paroxítonas. A distância só é marcada com /R/: 99% são oxítonas e 1% paroxítonas. No entanto, a grande maioria são verbos: das 19.205 oxítonas terminadas em /R/, 15.325 são verbos (80%) e como é sabido, interpretações mais fonéticas que fonológicas poderiam alterar significativamente esses dados, uma vez que /L/ pode ser interpretado como o glide [w] e o /R/ final dos infinitivos (ao menos dos verbos com maior freqüência de uso) não é realizado, aumentando, assim, o número de oxítonas terminadas em vogal. Formas não-lematizadas aumentariam os casos de paroxítonas com o final /S/.

Por mais que COLLISCHONN (1999:146) se esforce para mostrar que os procedimentos da consoante abstrata e da extrametricidade não são *ad hoc*, a análise dos dados mostra que não há sempre equivalências para a extrametricidade em outros registros da língua (como em ['arvi] para *árvore*). Também aqui, o estudo das freqüências de uso das palavras ajudará a resolver definitivamente esse ponto. Tampouco há, em muitos casos, qualquer tipo de realização dessas consoantes em formas derivadas, mas, pelo contrário, essas supostas "evidências" são, na verdade, resquícios de antigas analogias que subsistem muitas vezes como interfixos (MALKIEL 1964).

3. Estruturas silábicas do português

As informações deste capítulo se baseiam sobre o número total de sílabas que passa das seiscentas mil (634.320 BE; 626.324 na BES; 624.791 na BT; 616.477 na BS). A estrutura CV é tida como universal (CRYSTAL 2000:238), havendo línguas que só dispõem de duas estruturas: V e CV. No caso do português, essa estrutura não só está

freqüente, mas também é a mais numerosa, oscilando entre 65,6% (BS) e 67,3% (BE). Os casos GV (em que G é um *glide*) representam um número bastante pequeno: constantemente 0,3% em todas as bases.

Com relação à tonicidade, 55% das sílabas são pretônicas, 24% são tônicas, 18% são pós-tônicas finais e 3% são pós-tônicas não-finais. Em todas as posições, a estrutura CV é a mais freqüente conforme a tabela abaixo, referente à BT (as demais bases não geram diferenças significativas):

Tabela 4. Estruturas silábicas mais freqüentes do português conforme a tonicidade

silaba	Total		Tônicas		Pretônicas		Pós-tônicas não-finais		Pós-tônicas finais	
CV	378340	60.6%	70493	46.7%	203185	59.3%	16697	90.7%	87965	78.0%
CVC	96019	15.4%	44675	29.6%	47172	13.8%	8	0.0%	4164	3.7%
V	52592	8.4%	7981	5.3%	37365	10.9%	439	2.4%	6807	6.0%
CCV	27767	4.4%	5127	3.4%	19674	5.7%	1251	6.8%	1715	1.5%
VC	26826	4.3%	5663	3.8%	21112	6.2%	1	0.0%	50	0.0%
CGV	12200	2.0%	298	0.2%	1361	0.4%	8	0.0%	10533	9.3%
CVG	11453	1.8%	5884	3.9%	5568	1.6%	0	0.0%	1	0.0%
CVGC	6633	1.1%	6477	4.3%	121	0.0%	0	0.0%	35	0.0%
CCVC	5171	0.8%	1920	1.3%	3153	0.9%	0	0.0%	98	0.1%
GV	1828	0.3%	503	0.3%	413	0.1%	9	0.0%	903	0.8%
VG	1770	0.3%	429	0.3%	1339	0.4%	0	0.0%	2	0.0%
CCVG	909	0.1%	289	0.2%	620	0.2%	0	0.0%	0	0.0%
CVCC	772	0.1%	116	0.1%	547	0.2%	0	0.0%	109	0.1%
CGVC	608	0.1%	214	0.1%	271	0.1%	0	0.0%	123	0.1%
CCVCC	422	0.1%	82	0.1%	329	0.1%	0	0.0%	11	0.0%
VCC	310	0.0%	20	0.0%	290	0.1%	0	0.0%	0	0.0%
GVC	308	0.0%	271	0.2%	28	0.0%	0	0.0%	9	0.0%
VGC	334	0.0%	238	0.2%	96	0.0%	0	0.0%	0	0.0%
outras	529	0,1%	195	0,1%	121	0,0%	0	0,0%	213	0,2%
total	624791	100%	150875	100%	342765	100%	18413	100%	112738	100%

Como se analisa facilmente da tabela acima, as sílabas pós-tônicas são preponderantemente abertas (100% das não-finais e 97% das finais, sendo que as únicas exceções para as não-finais são: *ábaxe, antíspasto, cóferdã, oldfíeldia, pênalti, tsarévitche, fádingue, chálenger, lêmingue*). As sílabas abertas representam 67% das sílabas tônicas e 87% das pretônicas. As átonas, sobretudo as pós-tônicas, têm mais predileção por ataques complexos (7% nas pretônicas e pós-tônicas não-finais, 11% nas finais) do que as tônicas (5%). Isso fica evidente em CGV, que corresponde a 9% das sílabas pós-tônicas finais mas a 0% em outras posições.

4. As sílabas do português (variante paulista)

Numa interpretação segundo a variante paulista, o /N/ e o /L/ na coda não são considerados C, mas, respectivamente, como traço de vogal nasal e glide [w]. Há, ao todo, 54.019 sílabas terminadas em /N/ que se acrescentam às estruturas XV ou XVG.

Interpretando /L/ como glide o número de estruturas XVG aumenta, uma vez que há 12.024 sílabas nessa condição. A diferença pode ser vista nos totais da seguinte tabela:

Tabela 5. Resultados da diferença de interpretação de /N/ e /L/

	Padrão fonológico (/N/ e /L/ consoantes)		Interpretação paulista (vogais nasais e glide [w])	
CV	378340	60,6%	411576	65,9%
CVC	96019	15,4%	54117	8,7%
V	52590	8,4%	64011	10,2%
CCV	27767	4,4%	29694	4,8%
VC	26826	4,3%	12822	2,1%
CGV	12200	2,0%	12441	2,0%
CVG	11453	1,8%	26963	4,3%
CVGC	6633	1,1%	250	0,0%
CCVC	5171	0,8%	3438	0,6%
outras	7792	1,2%	9479	1,5%
Total	624791	100%	624791	100%

Na interpretação paulista de BT, as sílabas mais freqüentes são: [a] (5%), e [tʃi] (3%). Com 2% ocorrem as sílabas: [si], [ku], [ta], [ka], [du], [dʒi], [li]. Com 1%, a gama aumenta: [ɾa], [tu], [na], [ɾi], [ma], [mi], [o], [ni], [da], [la], [de], [te], [to], [e], [ko], [ɾu], [fi], [ẽ], [i], [pi], [pa], [bi], [nu], [mu], [ba], [zi], [ga], [sɐ̃w], [he], [es], [za], [se], [ʒi], [mẽ], [sa], [me], [mo], [lo]. Com 0,5%: [no], [pe], [lu], [ĩ], [po], [aɾ], [doɾ], [aw], [ki], [su], [kõ], [vi]. Com 0,4%: [le], [ɾo], [ne], [va], [ɐ̃], [fa], [bu], [tɾi], [so]. Com 0,3%: [gɾa], [kaɾ], [ʒe], [bo], [viw], [zu], [gu], [tɾa], [do], [ha], [fo], [u], [pu], [ɾju], [des], [taɾ], [ɾe], [sja], [fu], [dʒju]. Com 0,2%: [go], [be], [zaɾ], [zo], [ʃa], [kɐ̃], [tɾo], [vu], [laɾ], [tɔ], [pɾo], [ʒa], [zẽ], [ʎa], [fe], [ɲa], [ve], [sju], [sẽ], [ʒẽ], [ɐ], [mɐ̃], [ke], [ʃi], [dez], [hi], [lɔ], [kɾo], [peɾ], [pɾe], [lɐ̃], [ʒu], [tɾu], [naɾ], [ja], [ho] e assim por diante. Nas outras bases, não há mudanças de porcentagem, uma vez que a mudança de transcrição se dá nas sílabas menos freqüentes. $BE_{paulista}$ e $BES_{paulista}$, contudo, revelam aumentos apenas nos percentuais das labiais+[i] e nas velares+[i], sobretudo nas surdas. Desse modo [pi] oscila entre 0,8% ($BT_{paulista}$) e 1,2% ($BE_{paulista}$), [bi] respectivamente entre 0,8% e 0,9%, [dʒi] entre 1,6% e 1,7%, [ki] entre 0,5% e 1,1% e [gi] entre 0,1% e 1,1%. A explicação está, talvez, nas palavras prefixadas com *ab-, ob-, ad-* e outras palavras cultas gregas e latinas. Há entre 2600 a 3600 tipos de sílabas distintas.

É visível que, apesar de a sílaba mais freqüente ser do tipo V, a grande maioria é do tipo CV, surgindo, aos poucos, os casos com coda e, depois os com ataque complexo. Apesar de a lista ser extensa, não ocorre acima nenhum caso com a sílaba [ɛ]. A mais freqüente, [tɛ], aparece com cerca de 0,1%.

5. A sílaba CV em português (variante paulista)

Para as ocorrências de CV, os segmentos consonantais são, em ordem decrescente de freqüência: [k] (10,0%), [t] (9,4%), [m] (8,4%), [l] (7,5%), [s] (7,4%), [ɾ] (7,4%), [n] (7,1%), [d] (7,0%), [p] (5,1%), [tʃ] (4,7%), [b] (4,3%), [z] (3,8%),

[f] (3,3%), [ʒ] (2,6%), [g] (2,5%), [dʒ] (2,5%), [v] (2,4%), [h] (2,4%), [ʃ] (1,2%), [ɲ] (0,6%), [ʎ] (0,6%). Para as vogais temos: [i] (26,5%), [a] (24,1%), [u] (17,0%), [e] (10,6%), [o] (10,0%), [ẽ] (3,3%), [ɐ̃] (2,3%), [ɔ] (1,8%), [ɛ] (1,3%), [õ] (1,2%), [ĩ] (0,9%), [ɐ] (0,6%), [ũ] (0,5%). Os *glides* não estão sendo computados porque eles não fazem parte das CV. Os segmentos [a] e [ɐ] foram computados em um só fonema /a/, com 24,7%. Os segmentos [t] e [tʃ] também foram computados num único fonema /t/, com 14,1% (ultrapassando, assim, o primeiro lugar). Idem [d] e [dʒ] somados no fonema /d/ computam 9,5%, subindo, da 8ª posição para a 3ª. A seqüência de fonemas consonantais, da mais freqüente para a menos freqüente nas sílabas CV, é: t k d m l s ɾ n p b z f ʒ g v h ʃ ɲ ʎ. Curioso observar, nessa seqüência, que as consoantes [-voz] são preferidas às [+voz] quando há um par. A única exceção é o caso de [ʒ], que é preferida a [ʃ]. Os fonemas vocálicos continuam praticamente inalterados: i a u e o ẽ ɐ̃ ɔ ɛ õ ĩ ũ.

Por esses dados, é possível observar que com respeito ao ataque, metade das CV começa com /t/, /k/, /d/, /m/ ou /l/. Quanto à rima das CV, mais da metade termina com as vogais /i/ ou /a/. Por outro lado, todas as vogais nasais somadas juntamente com /ɔ/ e /ɛ/ equivalem a apenas 12% do total. É interessante constatar que os sons menos freqüentes – consonantais ou vocálicos – são, diacronicamente, mais recentes na língua, tendo, como ponto de partida, o latim. Abaixo de /b/, todas as consoantes são novas na língua, com exceção de /f/ e /g/ e, somadas, equivalem a 16,1% do total.

É de se perceber, contudo, que nem todas as vogais são igualmente produtivas. A CV mais freqüente para cada vogal oscila entre 13% a 20% das sílabas CV com essa mesma vogal, mas há exceções: /mẽ/ tem 28% de todas as CV com ẽ e /kõ/, com 58%. Abaixo de 1% estão seqüências com /fẽ/, /võ/, /vũ/, /vɔ/, /hu/, /nõ/, /gi/, /ge/, /gɛ/, /ʃẽ/, /ʃu/, /ʃõ/, /ʒõ/, /ʒɔ/, /ʎi/, /ʎe/, /ʎɛ/, /ʎẽ/, /ʎĩ/, /ʎu/, /ʎũ/, /ʎɔ/, /ɲi/, /ɲe/, /ɲɛ/, /ɲĩ/, /ɲõ/, /ɲũ/, /ɲɔ/. A seqüência /ʎõ/ não existe no *corpus*. Também /ʎĩ/ e /ʎũ/ ocorrem apenas uma vez cada, ou seja, tem produtividade praticamente nula.

O dado mais surpreendente foi que algumas C combinam melhor com algumas V do que com outras. Seguindo as seqüências decrescentes acima, supõe-se que, dada uma vogal V, a freqüência decresceria da seguinte forma: [t]+V, [k]+V, [d]+V etc. (conforme também decresce a freqüência da consoante). Tal fato não ocorre. Verificou-se, assim, a seqüência das consoantes com relação a cada vogal, ou seja, realizou-se uma *análise de posto* (ABDI, 2007) para determinar as preferências e rejeições dos acoplamentos C+V.

A Tabela 6 apresenta, para cada vogal, a seqüência de freqüências das correspondentes consoantes em ordem decrescente, ou seja, a ordem de *postos* das consoantes de cada vogal em sílabas CV. Assim, percebe-se que a sílaba CV com a vogal [i] mais freqüente é /ti/, seguida de /si/, /di/ etc até /ɲi/. Verifica-se na Tabela 6 que as ordens de freqüência das consoantes de algumas vogais são muito diferentes da observada quando se considera todo o conjunto das sílabas CV, apresentada na última linha.

Tabela 6 - Relação entre a ordem de preferência das consoantes em relação às vogais nas silabas CV do português (variante paulista)

Vogais	Consoantes																		
i	t	s	d	l	ɾ	m	n	f	p	b	z	ʒ	k	v	ʃ	h	g	ʎ	ɲ
ĩ	s	l	t	ɾ	p	z	ʃ	k	n	v	m	ʒ	d	f	b	h	g	ɲ	ʎ
e	d	t	h	s	m	p	l	n	ʒ	ɾ	b	f	v	k	z	ʃ	g	ʎ	ɲ
ẽ	m	z	s	ʒ	t	d	l	ɾ	n	p	v	b	h	k	g	ɲ	f	ʎ	ʃ
ɛ	t	s	n	m	l	p	ɾ	d	b	ʒ	f	h	k	v	z	ʃ	g	ʎ	ɲ
ɛ̃	k	m	l	t	ɾ	n	b	s	p	z	g	ʃ	d	v	h	f	ʒ	ɲ	ʎ
a	t	k	ɾ	m	n	d	l	p	g	b	z	s	v	f	h	ʃ	ʒ	ʎ	ɲ
ɔ	t	l	k	ɾ	n	p	m	b	f	d	s	g	z	ʃ	h	ʒ	v	ɲ	ʎ
o	t	k	m	l	n	p	ɾ	s	b	d	f	g	z	h	v	ʃ	ʒ	ɲ	ʎ
õ	k	p	m	d	l	t	b	g	h	s	ɾ	z	f	n	ʃ	v	ʒ	ɲ	ʎ
u	k	d	t	ɾ	n	m	l	s	b	z	g	p	f	v	ʒ	ɲ	ʃ	h	ʎ
ũ	k	f	m	ʒ	ɾ	n	b	l	t	g	d	p	ʃ	s	z	h	v	ɲ	ʎ
todas	t	k	d	m	l	s	ɾ	n	p	b	z	f	ʒ	g	v	h	ʃ	ɲ	ʎ

É flagrante observar aqui alguns detalhes: a velar surda [k] é predominante nos casos de vogais [+rec] e rejeitada com [-rec]. Nesses casos, a distância entre [k] e [g] diminui sensivelmente; por outro lado, [s] é mais freqüente com vogais [-rec] do que nas [+rec]. Impossível não associar esses dados com um fato diacrônico. Diacronicamente, a passagem de *[ki] > [si], *[ke] > [se] ocorreu no latim vulgar (provavelmente por meio de intermediários *[tsi], *[tse]). Trata-se de fato trivial, que se vê ao soletrarem-se as sílabas, no estilo das cartilhas: *ca, ce, ci, co, cu* ([ka], [se], [si], [ko], [ku]). Nosso *corpus,* no entanto, não é formado apenas de palavras provenientes do latim vulgar, mas, ao que tudo indica, as palavras que entraram para o léxico português, provindas do latim medieval e renascentista, pelo latim científico, pelo francês e por inúmeras outras línguas (inclusive não-indo-européias) parecem ter-se pautado pelo molde fônico primitivo e, paulatinamente, reforçado. A preferência de certas C com relação a certas V e vice-versa se vêem ainda em outros casos de forma menos contundente e tem casos aparentes de dissimilações (por exemplo: [v] rejeita sons arredondados como [õ], [ɔ], [o], [u], [ũ] e prefere não-arredondados como [ẽ], [e], [i]). Se NOGUEIRA (1958) tiver razão quando afirma que a assimilação é o princípio pelo qual se regem todas as transformações fonéticas, entendemos que os traços da consoante e os da vogal desempenham um grande papel na preferência de uma língua por uma sílaba CV em detrimento de outra. O antigo termo *assimilação* adquire, em teorias mais recentes, um papel muito importante, como, por exemplo, no modelo autossegmental, sobretudo no fenômeno do *espraiamento* (CLEMENTS & HUME, 1995) mas poderá ter grande aplicação, a partir de resultados como os nossos, em soluções derivadas do modelo da otimalidade, uma vez que há, como vimos, sílabas preferíveis (ótimas) e não só estruturas silábicas (PRINCE & SMOLENSKY, 2004). Em todo caso, trata-se de um campo novo a ser explorado.

Uma análise estatística mais completa foi empregada para quantificar as preferências e rejeições entre os acoplamentos de consoantes e vogais. Assim, estimamos a probabilidade média de cada vogal *v*, $p(v)$, por $p(v) = \frac{N_{\mathrm{CV}}(*, v)}{N_{\mathrm{CV}}(*, *)}$, onde $N_{CV}(*, v)$ é o número de sílabas CV formada por qualquer uma das consoantes e a

vogal *v*, e $N_{CV}(*,*)$ o número total de sílabas CV, no caso, 411.576. De maneira análoga, a probabilidade média de cada consoante *c* foi estimada por $p(c) = \frac{N_{CV}(c,*)}{N_{CV}(*,*)}$, e a probabilidade média de cada sílaba *cv* por $p(c,v) = \frac{N_{CV}(c,v)}{N_{CV}(*,*)}$.

Usando a definição de probabilidade condicional (MAGALHÃES & LIMA, 2002: 137), sabemos que $p(c,v) = p(c).p(v/c)$, onde $p(v/c)$ é a probabilidade de se encontrar a vogal *v* em uma sílaba cuja consoante é *c*. Escrevendo $p(v/c)$ em termos da probabilidade média da vogal *v* e de um fator de acoplamento, $\varphi(c,v)$, obtemos $p(v/c) = p(v).\varphi(c,v)$. A interpretação do fator de acoplamento $\varphi(c,v)$ é a seguinte: a ausência de preferência ou rejeição no acoplamento de *c* com *v*, implica que $\varphi(c,v)$ será igual a 1, indicando que a probabilidade da vogal *v* não se altera pela presença da consoante *c* e vice-versa. Nos casos de preferência, $\varphi(c,v)$ será maior do que 1, enquanto nas rejeições, $\varphi(c,v)$ será menor do que 1.

Manipulando as equações acima, é possível relacionar o fator de acoplamento com as probabilidades médias determinadas experimentalmente, resultando em

$$\varphi(c,v) = \frac{p(c,v)}{p(c).p(v)}. \qquad \text{(Equação 1)}$$

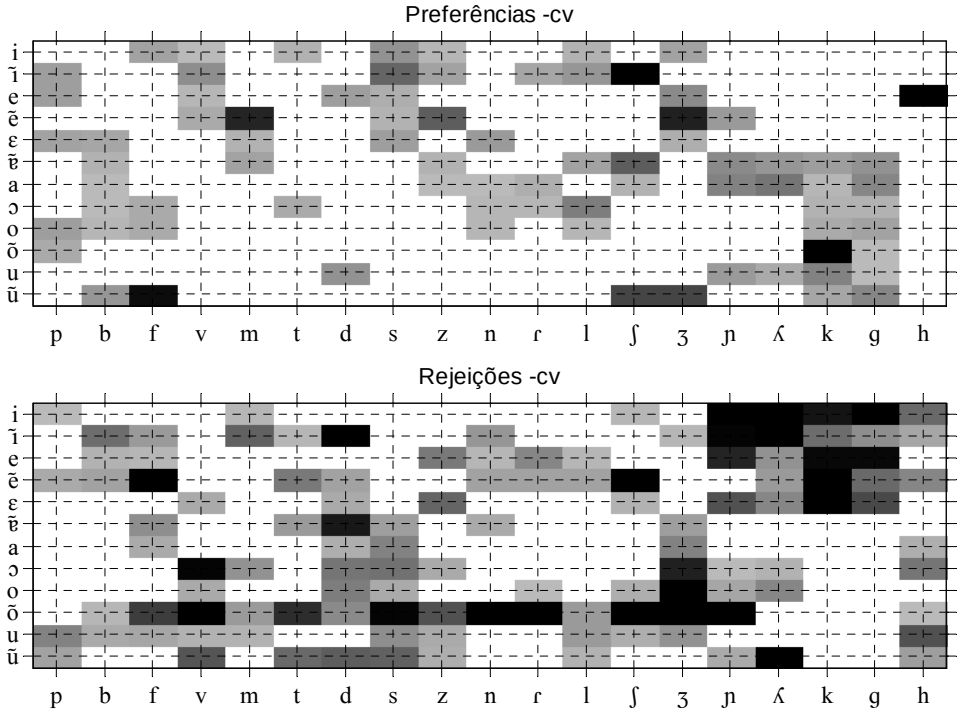

Figura 2 - Fatores de acoplamentos (Equação 1) superiores a 10% entre vogais e consoantes das sílabas CV do português (variante paulista).

Na Figura 2 são apresentados os acoplamentos e a intensidade das preferências e rejeições superiores a 10%. A análise da Figura 2 corrobora os resultados obtidos na análise de posto, porém, é muito mais rica em detalhes evidenciando, por exemplo, que as consoantes [k] e [g] possuem preferências e rejeições similares.

6. Conclusões

Nossos dados reforçam diversas concepções intuitivas ou obtidas em *corpora* bem menores (cf., por ex, BARBOSA, 1983: 223-229), que são pilares de muitas teorias, sobretudo a preponderância das paroxítonas e da sílaba CV na língua portuguesa, mas os dados particulares acerca das CV parecem absolutamente novos e dão um passo para entender aquilo que, durante séculos, os estudos literários e a terminologia não-científica (e, portanto, raramente adotada pela fonologia) tenta compreender com o nome de *eufonia.* A extensão do trabalho para outros tipos de estruturas silábicas, ou até mesmo para relações de preferência entre ataque, núcleo e coda parece promissora. Por exemplo, resultados preliminares mostraram que com relação às sílabas VC e VG, as vogais centrais e as posteriores combinam melhor com [ɾ] e [w] do que as anteriores, que combinam melhor com [s]. Dados de língua falada, que envolvam vocábulos flexionados (não-lematizados), associados com sua freqüência de uso dirão, em futuros artigos, se há de fato grande distinção entre dados da língua escrita e os de língua efetivamente falada. No entanto, os estudos de ALBANO *et alii* (1995), nos levam a prever que a diferença será, aparente, pequena.

7. Referências bibliográficas

ABDI, Hervé. The Kendall rank correlation. In: SALKIND, Neil (org.). *Encyclopedia of Measurement and Statistics*. Thousand Oaks (CA): Sage, 2007.

ALBANO, Eleonora *et alii.* Segment frequency and word structure in Brazilian Portuguese. *Proceedings of the XIVth IChS.* Estocolmo, V.3, p. 346-349, 1995.

BARBOSA, Jorge M. *Études de phonologie portugaise.* Évora: Universidade de Évora, 1983.

BISOL, Leda. O acento e o pé métrico binário. *Cadernos de estudos lingüísticos.* Campinas: Unicamp, v. 23, p. 83-101, 1992.

CLEMENTS, George N. & HUME, Elizabeth V. The internal organization of speech sounds. *In:* GOLDSMITH, John (org). *The handbook of phonological theory.* London: Blackwell, 1995.

COLLISCHONN, Gisela. O acento em português. *In:* BISOL, Leda (org.) *Introdução a estudos de fonologia do português brasileiro.* Porto Alegre: Edipucrs, 1999, p. 125-158.

CRYSTAL, David. *Dicionário de lingüística e fonética.* Rio de Janeiro: Zahar, 2000.

MALKIEL, Yakov. Generic analyses of word formation (1964) .*In:* SEBEOK, Thomas A. (org.) *Current trends in linguistics*. Paris: Mouton, 1970, p. 305-364.

MAGALHÃES, Marcos Nascimento & LIMA, Antonio Carlos Pedroso. *Noções de probabilidade e estatística.* São Paulo: Edusp, 2002.

NOGUEIRA, Rodrigo de Sá. *Tentativa de explicação dos fenómenos fonéticos em português.* Lisboa: Clássica, 1958.

PRINCE, Alan & SMOLENSKY, Paul. *Optimality theory: constraint interaction in generative grammar.* Malden/Oxford/Carlton: Blackwell, 2004.

SILVA, Adelaide H. P. *et alii.* Codificação fonológica informatizada do minidicionário Aurélio: um banco de dados para o estudo da fonologia portuguesa. *Anais XLI GEL.* Ribeirão Preto: UNESP, v. 1: 1321-7, 1993.

VIARO, Mário E. Relação entre produtividade e freqüência na produção do significado. *Estudos lingüísticos.* Campinas: Unicamp, v. 34: 1230-1235, 2005. Disponível em: www.gel.org.br/4publica-estudos2005-autores.htm