Journal of Data Science **8**(2010), 91-99

Language Rhythm Model Selection by Weighted Kappa

Viviana Giampaoli and Arnaldo Mandel
Universidade de São Paulo

Abstract: Given processes that assign binary vectors to data, one wish to test models that simulate those processes and uncover groupings in the processes. It is shown that a suitable test can be derived from a kappa type agreement measure. This is applied to analyze stress placement in spoken phrases, based on experimental data previously obtained. The processes were Portuguese speakers and the grouping corresponds to the Brazilian and European varieties of that language. Optimality Theory gave rise to different models. The agreement measure was successful in pointing the relative fitness of models to language varieties.

Key words: Kappa, language rhythm model.

1. Introduction

This work originates on the research for mathematical models for language rhythm. In particular, Portuguese is a language with two very distinct varieties, Brazilian Portuguese (BP) and European Portuguese (EP), which differ in many rhythmic aspects. One that stands out is the use of stress: when a person utters a phrase, the syllables that are stressed seem to reflect whether the person is a BP or EP speaker. A question posed in Sandalo *et al.* (2006) was whether such a stressing could be modelled within the confines of Optimality Theory (see Kager, 1999), in such a way that the provenience of a speaker could be gleaned by just looking at the stress pattern in speech; this article is, in a sense, a complement to that work.

In Optimality Theory models there are two main ingredients: *structures* and *restrictions*. The model has to choose among the structures; the restrictions are used to define a quasi-order on the structures and the choice is for the optimal structures (which may be many). For instance, from the restrictions one gets a real valued "cost" function on the structures, and the model chooses the minimal cost structures.

This is how optimality was instanced in that article: the structures associated to each phrase were factorizations of the phrase into segments of successive syllables. There were some feasibility criteria for the segments, in such a way that, from each such segmentation one could directly read a stress placement for the phrase. These segmentations can be conveniently encoded as paths on a directed graph, the *segmentation graph* of the phrase. Then, after some experimenting, a collection of restrictions was chosen; those are linguistically significant constructs and involve further information gathered from the phrase.

As a final step, there is a choice of weights for the restrictions. Each choice of restriction weights yields costs on the edges of each segmentation graph, and the preferred segmentations are those corresponding to shortest paths linking two special vertices.

In what follows, we refer to each weighting of the restrictions as a *model*. So, this is the data flow: from a phrase one gets the graph, from a model one gets costs in the graph, and then some paths. Those are decodified to produce a stress placement. In the end, a model produces for each phrase a collection of binary vectors, each describing a stress placement for the phrase.

This process was tested in the following experimental setting, as reported in Sandalo *et al.* (2006):

There was a fixed collection P of *phrases* for which different models could be tried. The phrases were given to speakers of both varieties of Portuguese (we further refer to them as *readers*); it was known, for each reader, which variety of Portuguese she speaks. They read the phrases aloud and the reading was recorded. The researchers then assigned to each reading a binary vector $O_r(p)$ (reader r, phrase p). As in the case of models, each binary vector associated to a phrase p has length equal to the number of syllables of p. The actual test bed consisted of 20 phrases and 4 readers.

It is worth noticing that in Portuguese each word has a *primary* stress, which does not vary; the variation occurs in the placement of the *secondary* stresses, which are needed for the utterance of long words. This has implications for the modelling, as will be explained in the next section.

An a priori grouping of the readers into two classes, BP and EP, was known. The main question was whether models could be chosen in such a way that this classification could be recovered, that is, whether one could choose, for each group, a model that reasonably predicted the stress placements uttered by its members. That was done in an ad-hoc manner, one model being chosen for each group and the adequacy of the models was argued in an intuitive manner.

We suggest here a quantitative approach for evaluating the models vis a vis the readers' grouping. That will be done through an *agreement measure*: members of the same group will have strong agreement within the group and with the certain

group model, while there will be little agreement between different groups and the another models. The (weighted) kappa coefficient (Cohen, 1960), which has been used to assess the degree of agreement between two ratings on presence or absence of a characteristic (see, for example, Fleiss, 1971, Poentius, 2000) turns out to be a useful statistics for this purpose. Applying those kappa-based criterion to the data and models of Sandalo et al (2006) results in a qualified vindication of those models: the two models proposed are each a good fit for one language variety and a poor fit for the other.

Given the small number of observations available, we reprocess the data through bootstrap techniques to enhance the confidence on the earlier results. Section 2 presents the techniques used and the results thus obtained. Section 3 present the bootstrapping. Section 4 contains concluding remarks.

2. Weighted Kappa

Cohen's kappa is an index of agreement of observations of categorical data. We will use it to measure the agreement between the observations of the readers and the vectors assigned by the models, for each input phrase. We present a short description of kappa, specialized for binary data and modified to allow for weights, following Poentius (2000).

We consider a binary vector as an assignment of category 0 or 1 to each of its components; in our application, the components are syllables, 0 means *not stressed* and 1 means *stressed*. Given a pair u, v of binary vectors of same length, the standard definition of kappa is based on the contingency table D of paired observations (u_k, v_k), where k ranges over the components. The weighted version allows for the presence of a nonnegative weight vector w of same length, so that each component k counts w_k for the contingency. More precisely, the 2×2 matrix D has entries $d_{ij} = \sum\{w_k : (u_k, v_k) = (i, j)\}$, where $i, j \in \{0, 1\}$. In particular, $d_{00} + d_{11}$ is the total weight of the components where the vectors agree. There is no loss of generality in supposing that $\sum_i w_i = 1$, and, since it simplifies some expressions, we assume it throughout. In particular, it follows that $d_{00} + d_{01} + d_{10} + d_{11} = 1$.

The marginal distributions of D give the weighted proportions of 0's and 1's in u and v. If, given these, the pairs (u_i, v_i) occurred independently, the expected agreement would be $P(u, v) = (d_{00} + d_{01})(d_{00} + d_{10}) + (d_{01} + d_{11})(d_{10} + d_{11})$. Kappa is defined based in the actual proportion of agreement, $A(u, v) = d_{00}+d_{11}$, centered and normalized relative to $P(u, v)$:

$$\kappa(u, v) = \frac{A(u, v) - P(u, v)}{1 - P(u, v)}.$$

It is easy to see that κ satisfies $-1 \leq \kappa(u, v) \leq 1$, the value 1 being attained

only if $u = v$, and -1 attained when they completely disagree, and the components where $u = 0$ and $v = 1$ have total weight $1/2$. The value 0 reflects independent observations.

We turn back now to the experimental setting, which consists of a collection P of *phrases*, and a set of *readers*. Each reader r assigns each phrase p a single binary vector $O_r(p)$. The agreement of phrase of readers will be greater in the most appropriate models for its variety of Portuguese; each *model* $\mathbf{m}$ assigns to each phrase p a nonempty set of binary vectors $O_{\mathbf{m}}(p)$. Recall that all vectors assigned to each phrase are the same length, the number of syllables.

We wish to recover the grouping of readers from agreement between binary vectors generated by phrases readers and those generated by models. So, for each phrase we compute the agreement between readers and models and summarize these values in order to drive the clustering decision .

For each phrase p, reader r and model $\mathbf{m}$, consider the following multiset:

$$K(r, \mathbf{m}, p) = \{\kappa_{puv} \,|\, \kappa_{puv} = \kappa(u, v), u = O_r(p),\ v \in O_{\mathbf{m}}(p)\}$$

and define $K(r, \mathbf{m})$ as the multiset union of $K(r, \mathbf{m}, p)$ over all $p \in P$.

We will consider two different weight types. Weighting 1 is uniform on the syllables, and is used as a ballpark measure. Weighting 2 is driven by a more accurate assessment of readers and models: linguistic reasons imply that for each phrase there are a few precisely identified coordinates in which all assigned vectors will agree [1]. It is natural then to assign weight 0 to those coordinates, and give equal weights to the others, keeping a total sum of 1. In what follows, all calculations will be done separately for each weighting.

The summary statistics of each $K(r, \mathbf{m})$ is presented in Table 1. Each column is labeled by reader (a, b, c, d), model $(\mathbf{b}, \mathbf{e})$, and weighting $(1, 2)$ (we use different symbols, instead of integer indices, in order to improve readability). The a priori grouping of readers was BP= $\{a, b\}$, EP= $\{c, d\}$.

Table 1: Summary statistics for $K(r, \mathbf{m})$

$r\mathbf{m}^{\mathrm{w}}$	$a\mathbf{b}^1$	$a\mathbf{b}^2$	$a\mathbf{e}^1$	$a\mathbf{e}^2$	$b\mathbf{b}^1$	$b\mathbf{b}^2$	$b\mathbf{e}^1$	$b\mathbf{e}^2$
min	0.350	-0.333	0.156	-0.333	0.444	0.000	0.156	-0.333
mean	0.793	0.592	0.634	0.270	0.822	0.664	0.625	0.254
median	0.792	0.607	0.675	0.315	0.798	0.650	0.675	0.315
max	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Std Dev	0.195	0.394	0.208	0.391	0.178	0.335	0.205	0.390
$r\mathbf{m}^{\mathrm{w}}$	$c\mathbf{b}^1$	$c\mathbf{b}^2$	$c\mathbf{e}^1$	$c\mathbf{e}^2$	$d\mathbf{b}^1$	$d\mathbf{b}^2$	$d\mathbf{e}^1$	$d\mathbf{e}^2$
min	0.333	-0.333	0.412	-0.207	0.125	-0.429	0.350	-0.333
mean	0.583	0.066	0.732	0.256	0.553	0.062	0.745	0.362
median	0.553	0.000	0.700	0.000	0.587	0.000	0.712	0.333
max	0.851	0.609	1.000	1.000	0.871	0.727	1.000	1.000
Std Dev	0.175	0.321	0.177	0.456	0.221	0.363	0.183	0.442

[1] For each word, the syllable with lexical accent gets a 1, and all succeeding syllables get a 0

First we noted that assessing what constitutes a good value for κ is problematic in itself and that different scales have been proposed. For example, Landis and Koch (1977) and Rietveld and van Hout (1993) consider $0.21 \leq \kappa \leq 0.40$ as indicating fair agreement, $0.40 \leq \kappa \leq 0.60$ as indicating moderate agreement, $0.61 \leq \kappa \leq 0.80$ and $0.81 \leq \kappa \leq 1.00$ as indicating substantial and almost perfect agreement, respectively. Krippendorff (1980), which discounts when $\kappa < 0.67$ and allows tentative conclusions when $0.67 \leq \kappa < 0.80$ and definite conclusions when $\kappa \geq 0.81$. In this work, we are interested in the comparison of the values of kappas in each case, thus here these scales serve as a guide and other studies would be necessary to determine the most appropriate scale for the weighted kappa. On a first glance one notices that, for each reader and model, weighting 2 affords kappa a smaller mean and bigger dispersion (Std Dev and SE mean) than weighting 1. That is to be expected, as the move from 1 to 2 was done by striking out components where agreement was fixed; as expected, this move accrued the discriminatory power of κ.

For each weighting, one notes that for readers a, b the mean κ is bigger for model **b** than for model **e**; the opposite occurs for readers c, d. That is the first evidence for our main conclusion about the data:

> Model **b** is a better fit for readers a and b, while model **be** is a better fit for readers c and d.

That was, indeed, the ad hoc conclusion offered in Sandalo *et al.* (2006); what we have shown is that their conclusion has a better support than intuition.

More support for this clustering is given by an analysis of how the adequacy of each model is evidenced at the individual phrase level. For this purpose for each reader, we consider the statistic $\Delta_r = \kappa_{\mathbf{e}} - \kappa_{\mathbf{b}}$, where $\kappa_{\mathbf{e}} \in Kr\mathbf{e}p$, $\kappa_{\mathbf{b}} \in Kr\mathbf{b}p$, and p ranges over all phrases. We expect Δ_r to be negative for $r = a, b$, because the agreement measure $\kappa_{\mathbf{e}}$ would be less than $\kappa_{\mathbf{b}}$ for $r \in$ BP, and for it to be positive for $r = c, d$.

The summary statistics for these differences Δ_r, over all phrases, are presented in Table 2, columns indexed by reader and weighting [2].

Table 2: Summary statistics for Δ_r

r^{w}	a^1	a^2	b^1	b^2	c^1	c^2	d^1	d^2
min	-0.844	-1.333	-0.844	-1.333	-0.294	-0.771	-0.325	-0.901
mean	-0.155	-0.331	-0.190	-0.412	0.158	0.188	0.222	0.320
median	-0.143	-0.400	-0.181	-0.458	0.136	0.028	0.228	0.235
max	0.650	1.333	0.556	0.917	0.667	1.333	0.833	1.375
Std Dev	0.304	0.582	0.287	0.540	0.209	0.526	0.251	0.565

[2]Note that the weighting is denoted by a superscript. In the following sections, whenever we want to make explicit the dependence of any statistic on weights, we use a superscript w.

These differences are generally bigger when weighting 2 is used, again attesting for its better discriminating ability.

3. Bootstrap Based Inference

The reason for using bootstrap inference is that hypothesis tests and confidence intervals based on asymptotic theory can be seriously misleading when the sample size is not large. Here we use bootstrap to evaluate the confidence intervals for δ_r^w, the statistical mean of Δ_r^w, for each reader r and weighting w. Note that for each reader the set of kappa values is naturally stratified by input phrases, and each stratum is correlated from inception. As we do not have any hypothesis or knowledge on the theoretical distribution of kappa, we appeal to non-parametric methods.

For this reason we consider non-parametric bootstrap confidence limits and the achieved significance level (ASL) of the test for the comparison of kappas (see, for example, Efron and Tibshirani, 1993).

For each resample, a bootstrap sample is drawn separately for each stratum $\{\kappa_{\mathbf{e}} - \kappa_{\mathbf{b}} \,|\, \kappa_{\mathbf{e}} \in Kr\mathbf{e}p, \kappa_{\mathbf{b}} \in Kr\mathbf{b}p\}$, and those are combined to give the full resample. The sample mean $\bar{\Delta}_r^w$ is calculated for the resample as a whole. The bootstrap summary statistics based on 10,000 bootstrap replications are presented in Table 3. Empirical percentiles and BCa (Bias-corrected accelerated) confidence limits are shown in Table 4.

Table 3: Bootstrap Summary Statistics

$\bar{\Delta}_r^{\mathrm{w}}$	a^1	a^2	b^1	b^2	c^1	c^2	d^1	d^2
Observed	-0.155	-0.331	-0.190	-0.412	0.158	0.188	0.222	0.320
Bias	0.000	-0.000	-0.000	0.001	-0.000	-0.000	-0.000	-0.000
Mean	-0.155	-0.331	-0.190	-0.412	0.158	0.188	0.221	0.320
SE	0.020	0.042	0.020	0.042	0.017	0.044	0.019	0.047

Table 4: Empirical percentiles and BCa confidence limits based on 10,000 bootstrap replications

	Empirical Percentiles				BCa Confidence Limits			
Δ_r^{w}	2.5 %	5%	95%	97.5%	2.5%	5%	95%	97.5%
a^1	-0.194	-0.188	-0.122	-0.116	-0.195	-0.188	-0.123	-0.116
a^2	-0.416	-0.402	-0.262	-0.248	-0.416	-0.402	-0.262	-0.248
b^1	-0.230	-0.224	-0.157	-0.150	-0.230	-0.224	-0.157	-0.150
b^2	-0.493	-0.481	-0.343	-0.330	-0.495	-0.482	-0.344	-0.331
c^1	0.125	0.130	0.186	0.192	0.126	0.131	0.187	0.192
c^2	0.102	0.115	0.261	0.274	0.103	0.116	0.262	0.275
d^1	0.184	0.190	0.253	0.258	0.185	0.191	0.254	0.259
d^2	0.227	0.241	0.397	0.413	0.228	0.243	0.398	0.413

We can observe on Tables 3 and 4 that none of the intervals contains the zero value, thus the previous conclusions about model-reader fit are confirmed.

If the BCa method is considered, we could claim that the null hypothesis $\delta_r^w = 0$ is rejected at the 0.05 level, for all r. The bootstrap ASL testing the same null hypothesis $\delta_r^w = 0$ considers

$$\widehat{ASL} = \frac{\#\{b|\bar{\Delta}_r^w(b) < 0\}}{B}.$$

i.e., the proportion of bootstrap replications (b) less than zero, with $B = 10,000$. The minimum and maximum observed values in the 10,000 bootstrap replications of $\delta_r^w = 0$ are negative for the readers a and b and positive for c and d. These results reinforce the above-mentioned conclusion on the relationship model-reader.

Table 5: Summary Statistics in the contaminated sets

$\kappa_{\mathbf{m}}^w$		$a\mathbf{b}^1$	$a\mathbf{b}^2$	$a\mathbf{e}^1$	$a\mathbf{e}^2$
$p = 1$	min	0.350	-0.333	0.156	-0.333
	mean	0.793	0.592	0.634	0.270
	median	0.792	0.607	0.675	0.315
	max	1.000	1.000	1.000	1.000
	Std Dev	0.195	0.394	0.208	0.391
$p = 5$	min	0.350	-0.333	0.156	-0.333
	mean	0.748	0.483	0.645	0.270
	median	0.732	0.524	0.678	0.265
	max	1.000	1.000	1.000	1.000
	Std Dev	0.196	0.408	0.210	0.399
$p = 10$	min	0.125	-0.428	0.350	-0.333
	mean	0.645	0.281	0.710	0.343
	median	0.675	0.333	0.700	0.333
	max	1.000	1.000	1.000	1.000
	Std Dev	0.245	0.462	0.183	0.418
$p = 15$	min	0.125	-0.429	0.329	-0.333
	mean	0.641	0.273	0.723	0.351
	median	0.680	0.233	0.704	0.333
	max	1.000	1.000	1.000	1.000
	Std Dev	0.259	0.474	0.188	0.426

3.1 Sensitivity of kappa

We can ask what is the performance of kappa for a mix of phrases for two readers of different groups; if kappa were not sensitive to this mix, its reliability would be questionable. To answer this, we randomly selected a reader of each group, a and d. In the set of phrases of reader a we randomly replaced $p = 1, 5, 10, 15$ phrases for phrases of reader d, constructing so called *contaminated*

data sets. In Table 5, we present the value of kappa associated to reader a for model $\mathbf{m}$ and weight w ($\kappa^w_{\mathbf{m}}$); we observe that $\kappa^w_{\mathbf{b}}$ decreases and $\kappa^w_{\mathbf{e}}$ increases with the increase in the number of substituted phrases. Bootstrap Summary Statistics, Empirical Percentiles and BCa Confidence Limits for the statistics mean $\Delta^w_a = \kappa^w_{\mathbf{e}} - \kappa^w_{\mathbf{b}}$ were obtained. We observe that the values of Δ^w_a tend to be positive when the number of substituted phrases increases, being practically null when the number of phrases of the two languages is the same. All these results are expected indicating a very good performance of kappa.

3.2 Selection of the weights

One can argue that, for a given reader r. the higher the absolute value of Δ^w_r, the bigger the evidence that one of the two models fits r. We compare now the two weightings on this basis, by studying the statistics $D_r = |\Delta^2_r| - |\Delta^1_r|$, where superscripts again indicate weightings.

We obtained the usual summary statistics, bootstrap summary statistics, bootstrap confidence limits and minimum and maximum values of the replicate bootstrap of $D_r > 0$ for each reader . These results confirm, as expected, there is strong evidence that $D_r > 0$, a further support to the intuition that weighting 2 is a better choice than weighting 1.

4. Conclusion

We have shown an example where weighted kappa can be a useful agreement measure for model selection. The use of stratified bootstrap was driven by the small sample size, and by the multi-valued character of the models. The analysis also exemplifies that a judicious choice of weighting can lead to more supported conclusions.

This can be further improved: given the quantitative quality measure given by kappa, one could aim to eliminate the ad-hoc component in the choice of models. Such a choice can perhaps be construed as an optimization problem in a suitable "model space".

Acknowledgements

This work received partial financial support from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP): PRONEX/FAPESP's Project Stochastic behaviour, critical phenomena and rhythmic pattern identification in natural languages (grant number 03/09930-9). The authors would like to thank Professors Ricardo Fraiman, Alejandro Frery and Antonio Galves for their enlightening sug-

gestions. Special thanks are due to the referee for the careful reading and deep comments.

References

Cohen, J. (1960). A coefficient of agreement for nominal scale. *Educational and Psychological Measurement* **20**, 37-46.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman and Hall.

Fleiss, J. L. (1971). *Statistical Methods for Rates and Proportions*. Second Edition. John Wiley.

Kager, René. (1999). *Optimality Theory*. Cambridge University Press.

Krippendorff, Klaus. (1980). *Content Analysis: An Introduction to Its Methodology*. Sage Publications.

Landis, J. e Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, **33**, 159-174.

Poentius, R. J. Jr. (2000). Quantification error versus location error in comparison of categorical maps. *Photogrammetric Engineering and Remote Sensing* **66**, 1011-1016.

Rietveld, T. and R. van Hout (1993). *Statistical Techniques for the Study of Language and Language Behaviour*. Mouton de Gruyter.

Sandalo, F., Abaurre, M. B., Mandel, A., Galves, C. (2006), The Sotaq optimality based computer program and secondary stress in two varieties of Portuguese, *Probus* **18**, 97-125.

Received March 4, 2008; accepted July 13, 2008.

Viviana Giampaoli
Departamento de Estatística
Universidade de São Paulo
CEP 05315-970
São Paulo, Brasil
vivig@ime.usp.br

Arnaldo Mandel
Departamento de Ciência da Computação
Universidade de São Paulo
CEP 05315-970
São Paulo, Brasil
am@ime.usp.br