Kappa statistic for judgement agreement in Sociolinguistic

Raquel Meister Ko. Freitag[[1]](#footnote-1)

ABSTRACT: Perception studies have required the development of new techniques, as well as new ways of analyzing the data. In this text, a proposal for the measurement of the agreement of judgments in perception tests is discussed, the Kappa statistic. The analysis of a subjective reaction test of the variants /t, d/ considering aesthetic, rhythmic and dialectal factors with Kappa-statistic shows aspects of the inter-rater agreement and reliability that can explain how the linguistic change is driven in the community. The analysis also allows test in a perception study the sex/gender bias observed in production. The results sign that Kappa-statistics is a tool that can improve the power explanatory of perceptual studies in Sociolinguistic.

KEYWORDS: Perception study; statistic tools; agreement; palatalization.

RESUMO: Estudos de percepção têm demandado o desenvolvimento de novas técnicas observacionais e experimentais, assim como novas formas de tratar os dados obtidos. Neste texto, é discutida uma proposta para a aferição da força de convergência de julgamentos de testes de percepção, por meio do teste Kappa. A análise de um teste de reação subjetiva sobre o uso das variantes /t,d/ considerando fatores estéticos, rítmicos e dialetais com a estatística Kappa mostra aspectos da concordância e da confiabilidade entre os avaliadores que podem ser usados para explicar como a mudança linguística se dá na comunidade. A análise também permite identificar em um estudo da percepção o viés de sexo/gênero observado nos estudos de produção. Os resultados apontam que a estatística Kappa é uma ferramenta que pode ampliar o poder explanatório dos estudos de percepção em Sociolinguística.

PALAVRAS-CHAVE: Estudo de percepção; ferramentas estatísticas; concordância de juízes; palatalização.

**Introduction**[[2]](#footnote-2)

Subjective reaction tests were presented since the first approaches of Sociolinguistic, but in the production studies researchers have paying attention to the methods of data collection in order to resolve observer’s paradox, and, by consequence, this type of datum leaved to reach the development of an (the most) appropriate mathematical model to the quantitative treatment. The same cannot be said about the perception studies.

Attending the call for papers of Relin about as new ways of analyzing perception in Sociolinguistic, this paper provides a proposal for a new look to the quantitative approach of datum collected in perception studies in Sociolinguistic, as happened with the first studies of production in Sociolinguistic. Methodological discussions about how is possible get more statistical significance to quantitative approach in Sociolinguistic have been held since the first studies, for example, the quantitative approach to variation of /R/ in department stores in New York city, which initially was explored only in percentage (LABOV, 1972, 2006), and posteriorly in logistic regression with Varbrul (PAULILLO, 2002), or it is used to compare logistic regression and decision trees methods (EDDINGTON, 2010; TAGLIAMONTE; BAAYEN, 2012). Each one of these approaches with the same dataset had contributed to design the pattern of analysis adopted nowadays for data from production studies in Sociolinguistic. But there is still no consistency about the most appropriate mathematical model to deal with data from perceptual studies and the most common is only the percentage, like in the pioneering production studies.

This text aims 1) to explore the concepts of reliability and agreement in judgement tests (as in subjective reaction test, for example), considering the inherent inter-rates and intra-rates variance, 2) to present a mathematical model to quantifying the variability in judgement tests, the Kappa-statistic, and 3) to reanalyze the previous dataset using Kappa-statistics, in order to improve the perceptual approaches in Sociolinguistic.

**Quantitative approach to perceptual studies in Sociolinguistics**

Perception studies in Sociolinguistic try to answer questions about the role of speakers which presupposes to infer a certain direction towards to change by the data. When the speakers answer questions as “Does the speech of this person sound “ugly” or “beautiful”?”, “Does the speech of this person sound “quick” or “slow”?”, the account of answers not necessarily is consensual: if all the speakers answered “quick”, or “ugly”, there is a consensus, and it is a completed change. In incoming change processes, the variability in answers is expected and desirable; and identifying the degree of agreement in the answers can help to interprete as a direction of the linguistic change. It allows to identify the sex/gender, dialectal, educational, or other social or cultural bias driving the variant choices. For example, women (or educated, or urban people) can attribute more positive values to one variant than the men (or non-educated, or rural people).

One technique to account it is the percentage of judgements as measure of agreement. It is apparently easy to calculate it and to interpret it, because it consists in the compute of occurrences to each feature rated by judges. But the compute of percentage does not account for the agreement that is consequence from random variability or chance. Fieldworks that study the effects of pattern of judgement, as psychology, education, medicine, etc. have adopted Kappa statistics to consider the effect of random variability; and this paper propose to broad this scope to covert also the perception studies in Sociolinguistic. The concepts of reliability and agreement in psychometrics are explained towards to detail Kappa statistic and its applications in perceptual approaches in Sociolinguistic.

***Reliability and agreement***

The degree of agreement among speaker judges can be measured considering intra and inter rater pattern, which involves the concepts of **reliability** and **agreement**. Reliability concerns to the relatively consistency of a measuring, whether a test, a scale or a pattern are consistent. It is different of agreement, which refers to the convergence in the results.

This account considers the intra-rater (the consistency of rates, or the degree of agreement shown by the same rater at a distance of time) and the inter-rater reliability (the relative consistency in ratings, or the degree of agreement between the choices made by two or more independent judges) (LEBRETON; SENTER, 2008). It means, whether judges rank order targets in a manner that is relatively consistent with yourself and with other judges, not only in scores, but rather with the equivalence. The inter-rater agreement refers to the absolute consensus in scores provided by multiple judges for one or more targets, and intra-rater agreement refers to the absolute consensus in scores from the same rater at a distance of time.

Raters in this approach refer to judges, annotators, interviewers, transcribers (or anyone person or entity in action, like anesthesiologists, psychiatrists, nurses, etc.) and the subject of the rating can be persons, things, processes, outcomes, datum (time since of occurrence of target behavior, for example) (TINSLY; WEISS, 1975, 2000).

Measuring reliability and agreement in perception studies can help to outline the effects of demographic differences between groups of raters (inter-rater reliability and agreement), and the consistency of raters in their own judgments (intra-rater reliability and agreement). And the Kappa statistics is a measure to show and test reliability among multiple raters for categorical data, a common situation in subjective reaction tests.

***The Kappa-statistics***

Kappa (κ) is a statistical coefficient that measures the degree of accuracy and reliability between two raters who classify each one subjects in a rating scale.

This first version of Kappa coefficient was introduced by Cohen (1960), considering only two raters and a nominal scale; another version of Kappa enlarged the number of judges (Fleiss’s Kappa) and the type of scale (centered-weight Kappa), as seen above. The coefficient expansion is labeled Kappa-like statistics or Kappa-statistics (POSNER et al, 1990).

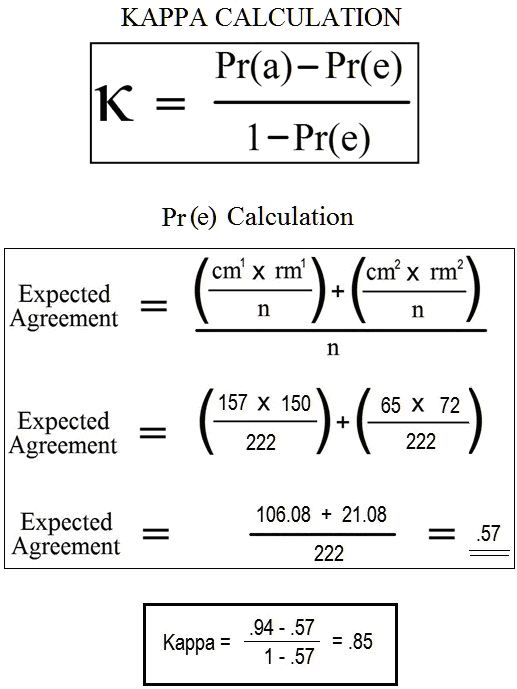
While Cohen’s Kappa is an agreement measure based on the pattern of two raters in a categorical scale for the subjects, Fleiss’s Kappa is applied among more than two fixed or random raters. Expanding the scope, weighted Kappa (FLEISS; COHEN, 1973) is an agreement measure to ordinal data (CHOUDHARY; NAGAJARA, 2017). So, the choice among specific Kappa test depends on how many raters and whether the ratings are nominal or ordinal (figure 1).

**Figure 1:** Kappa-statistics application.

|  |  |  |
| --- | --- | --- |
|  | *Scale measure* | |
| *Number of raters* | **nominal/categorical** | **ordinal** |
| two raters | Cohen’s Kappa | Weighted Kappa |
| + two raters | Fleiss’s Kappa |  |

Based on Cohen (1960), the basic assumption of Kappa-statistics is the ratio of the difference between the expected Pr (e) and observed agreement Pr(a), as in (1).

(1)



The Kappa coefficient ranges from -1 to +1, where 0 represents the amount of agreement expected from random chance, and 1 represents perfect agreement between the raters (figure 2).

**Figure 2:** Value of Kappa and its interpretation (LANDIS; KOCH, 1977, p. 165)

|  |  |
| --- | --- |
| Value of Kappa | Level of agreement |
| 0.00 – 0.20 | None |
| 0.21 – 0.39 | Minimal |
| 0.40 – 0.59 | Weak |
| 0.60 – 0.79 | Moderate |
| 0.80 – 0.90 | Strong |
| Above 0.90 | Almost perfect |

It is possible that Kappa is negative, and it means that the two raters agreed less than would be expected just by chance (and it is particularly important in Sociolinguistic, as seen below).

Fleiss’s Kappa (FLEISS, 1971) provides a measure of agreement among 3 or more raters; the formula is the same of (1), but in Fleiss’s Kappa Pr(a) is the observed proportion of the pairwise agreement among the m trials and Pr(e) is the expected proportion of agreement if the ratings from one trial is independent of another.

Some aspects must be considered in the adoption of Kappa-statistic: 1) The rating is measured on a nominal scale, with ordinal or nominal variables and the response categories are mutually exclusive (no categories overlap); 2) The rating is paired observations of the same phenomenon, it means all the raters assess the same observations; 3) The raters are independent, it means one rater's judgement does not affect the others rater's judgement.

Kappa-statistic has several applications in research fields where judgement tasks are required, and used it to assess agreement between classifications made on the same participants on different occasions, between classifications made by different observers or between classifications made by different methods. Aman and Szpavowicz (2007) used Coehn’s Kappa to measure the consensus agreement between two pairwise of judgments which consists in identifying emotions in texts based on an appraisal framework, that includes attitudes, judgments and emotions. The procedure consisted in to compare the results of one judgement, the specialist which stablished the golden pattern, against the others three, naive, and to calculate the mean of agreement with golden pattern in labeling emotion/non-emotion, emotion categories and emotion intensities. In this famous study about the universality of facial expressions, Paul Ekman and collaborators (1987) adopted Kappa-statistic in a judgement test which consists in the selection of only one emotion term for each expression, in a set of three expression for each 10 emotion (180 subjects) by 10 judgement for each one of 10 cultures.

Carletta (1996, p. 253) points the interest of computational linguistics and cognitive science in subjective judgements and claims the Kappa-statistic is as a uniform measure of reliability: “Kappa is widely accepted in the field of content analysis. It is interpretable, allows different results to be compared, and suggests a set of diagnostics in cases where the reliability results are not good enough for the required purpose. We suggest that this measure be adopted more widely within our own research community.”

For perceptual studies in Sociolinguistic approach, reliability can be considered as the ratio of true score variance to total variance and Kappa-statistic measures pairwise agreement among a set of raters making category judgments, correcting for expected chance agreement; the result is affected by skewed distribution of categories and by degree of disagree between the raters. Next section provides a hypothetical example of inter-rater agreement in the discrimination of sounds and how Kappa-statistic can elucidate this question.

***Explaining the Kappa-statistics***

In the phonetics transcription process, continuous sounds are cut off impressionistically; in some contexts, as /t/ and /d/, it is relatively clear; but there are certain contexts where the cut off is subjective, as one results from palatalization process, that produces a gradient of sounds, for example between /t/ and /tʃ/ in Brazilian Portuguese (FREITAG; SOUZA, 2016).

Suppose a task in which two transcribers were rated a set of 40 sounds and they must rate each one them in /t/ or /tʃ/. There is no gold standard; it means the rates are not right or wrong; the objective is identifying the agreement between the transcribers deciding if a sound is /t/ or /tʃ/ and the result based on these rates can provide a gold standard for validating next transcriptions. To constitute a dataset, the judge ratings (observation) are put in the rows of a table, and the stimuli are put in the columns, as in the spreadsheets 1 and 2 in Figure 3.

**Figure 3:** Spreadsheets of transcribers agreement.



The ideal scenario is the perfect agreement which is when transcriber 1 and transcriber 2 agree in the same judgement for all the 40 sounds, as seen in the column “agreement” in the spreadsheet 1.

Both the transcribers agree that 25 sounds from the dataset are alveolar stops and 15 are palatals. Based on the judgements of these transcribers, in the dataset, the rate of palatalization is 63%, and 38% for alveolar stops. And both transcribers totally agree this (100% of agreement). This is an ideal and unlikely scenario, as the next one, from spreadsheet 2, where the distribution of the rates is equitable; transcriber 1 and 2 rating 50/50 of the sounds randomly in t or tʃ, and the rate of palatalization is the same than the alveolar stops, 50% of cases, but, the transcribers agree only in 25% of cases.

If the task were to select a transcriber there is no doubt that the transcribers in the case of spreadsheet 1 are better than the transcribers in the case of spreadsheet 2; in the case of spreadsheet 2, it must be identified what is the transcriber in disagreement. It is not possible to identify based only with percent agreement. To do it, it must be to add one more rater and to compare its mean of agreement with the others one and thus deciding what is the transcriber in disagreement. It requires the Kappa-statistic. The first step is transforming the data in the spreadsheets in *n* x *n* contingency table, as the layout in the Figure 4:

**Figure 4:** Layout of contingency table for Kappa statistics.



The second step is proceeding the calculations for the expected agreement Pr(e), and observed agreement Pr(a). The observed agreement is calculated by the summing the frequencies in the main diagonal cells (cells a and d) dividing by n, as in (2).

(2)

Pr(a) = (a + d)/n

Multiplying by 100, it is the percentage agreement, and it is one measure adopted as measure of agreement between raters.

The proportion of expected agreement is based on the assumption that the ratings are independent between transcribers. Therefore, the frequency of chance agreement for a sound be /t/ or /tʃ/ is calculated by multiplying the marginal totals corresponding to each cell on the main diagonal and dividing by n. And, the proportion of expected agreement is calculated by summing across chance agreement in these cells and dividing by n, as in (3).

(3)

Pr(e) = [(f1\*g1)/n +(g2\*f2)/n)]/n

And, in third step, these results are applying in Kappa’s formula, in (4):

(4)

κ = 1 – (1 - Pr(a))/(1- Pr(e))

Resuming the spreadsheets 1 and 2, the sum of ratings is exposed in figure 5; the Kappa coefficient points to two different scenarios: spreadsheet 1 from figure 3 points to the perfect agreement, and spreadsheet 2 in figure 3 it points to the randomly agreement.

**Figure 5:** Summarizing spreadsheets of agreement.



As it seems obvious, perfect agreement has highest Kappa coefficient and the randomly agreement is 0.0, in the level of none agreement, according Landis and Koch (1977) guidelines presented at figure 2. But not always the results are so obvious as these. Suppose other two scenarios, A and B (figure 6): both present 90% agreement. But Kappa coefficient suggests that in scenario B the transcribers present strong agreement, with κ = 0.80, and in A, moderate agreement, κ = 0.60. Comparing the results in order to establish the gold standard in impressionistic transcription of /t,tʃ/, the scenario B is more reliable than the scenario A.

Considering now the scenarios C and D: in scenario C, percent agreement is 60% and κ = 0.20, suggesting none agreement; in D, percent agreement is 40% and κ = -0.20. McHugh (2012) adverts that Kappa coefficient above 0 is an indicator of a serious problem in research data, or, in this case, in the transcription process (it can happen if one of the transcribers is a specialist and another one is a naïve, for example). A negative Kappa signs that agreement is worse than expected, or disagreement; and a large negative Kappa represents great disagreement between the raters; it is very bad for clinical studies, as it is the most common application of Kappa statistics. But, it is particularly interesting for perception studies in Sociolinguistic (considering that any agreement less than 1.00 is also a measure of the disagreement between the raters), in order to allow evidence for inter rater bias and the strength of agreement with the measure of the extent to which raters assign the same score to the same variable.

**Figure 6:** Four scenarios for agreement transcribers.



Kappa results can use to test rater independence (testing the null hypothesis that there is no more agreement than might occur by chance), and to quantify the level of agreement, and it can provide evidences whether judges’s criteria of a perceptual study has been used consistently.

**Reanalyzing a perceptual study with Kappa-statistic**

Looking for the goals of this paper, a new analysis from the dataset of other study was carried out about. The research question concerns the perception how do /t/ and /d/ sound in Sergipe: palatal or alveolar stop. Palatalization of /t/ and /d/ followed by the vowel /i/, like in “tia” and “dia”, is called regressive palatalization; there is another process in Brazilian Portuguese, when the /t/ and /d/ is preceded by the glide /y/, and is called progressive palatalization, like in “oito” and “peito”.

The standard of urban dialects almost all over Brazil, in regressive contexts, is the palatal /t/ and /d/ realization: where the alveolar stop realization is more productive it is related to certain dialectal scopes and certain indexical fields. In the dialectal scope of the South of Brazil, alveolar stop realization is related to Italian immigrant descendants (BATTISTI et al., 2007, BATTISTI; DORNELLES, 2015); borders with Hispanic countries, like Argentina and Uruguay (CARVALHO, 2004; CASTAÑEDA, 2016); or immigrant descendants in general (BISOL, 1991). In the dialectal scope of Southeastern Brazil, alveolar stop realization is related to: “caipira” and “nordestino”, people who came from the Northeastern region (OUSHIRO, 2016). In all the cases, in both regions, the value associated to the alveolar stop realization is negative. On other hand, progressive palatalization is less recurrent, more restricted to certain dialectal areas, like countryside of Sergipe, and highly stigmatized (MOTA, 2008, FREITAG, 2015).

Production studies results in Sergipe suggest a change in progress, but these studies don’t inform the social forces driving this process nor how the innovative variant is evaluated by community.

Freitag e Santos (2016) presented results of the perception of undergraduates about the variation in /t,d/ realization in Brazilian Portuguese, in a dialectal area where the change is incoming, in Sergipe, Brazil. This previous analysis was strictly exploratory and considered only the percent of responses, without anyone statistical treatment for the data.

The subjects of rating composed a verbal guise, which is a subjective reaction test with stimuli collected in the sociolinguistic interviews from Falares Sergipanos database (FREITAG, 2013), in conditions approximate to minimal pairs of isolated words (LADEGAARD, 2000; DAILEY; GILES; JANSMA, 2005), without manipulation.

The task consisted in the judges hear the stimuli to respond a sequence of questions regarding aesthetical, rhythm and regional features about the speak way, follows Cardoso (2015): “Does the speech of this person sound “ugly” or “beautiful”?”, “Does the speech of this person sound “quick” or “slow”?” The response to the verbal guise is binary, in terms of opposite values, like “ugly” or “beautiful”, “quick” or “slow”.

The raters were 36 volunteer undergraduates at Federal University of Sergipe born and live on the capital and surround urban region, stratified equally in male and female.The rating task was run in PsychoPy v.1.82.01 (PEIRCE, 2007), in the Laboratório Multiusuário de Informática e Documentação (Lamid) room, at Federal University of Sergipe.

The new analysis intents to expand the discussion for beyond the percentage of responses, measuring the agreement among the raters and the converge of ratings among the sample: is the percentage of ratings in accord than the expected agreement? And is the observed agreement convergent with the percentage of ratings? As explained, the sample is relatively homogeneous, composing by undergraduates (same educational level and presumably the same age group). Only the sex/gender of raters was controlled: as in the production studies of palatalization process, does sex/gender bias affect the perception results? These questions guide the analysis.

There are 36 raters, each one rated all stimuli one times, independently, in a categorical scale, thus, Fleiss’s Kappa is the measure to be adopted. The first step to carry out a kappa-statistics analysis is preparing the dataset. The spreadsheet of ratings can be summarized in a 3-way table matrix, one for each feature rated. The second step is the account of observed and expected frequencies to calculate the Fleiss’s Kappa. It can make with a spreadsheet software, like ®Microsoft Excel, or a statistical software, like ®SPSS or R (R CORE TEAM, 2018). In last one, Fleiss’s Kappa was calculated with irr package, kappam.fleiss function (GAMER, 2016), but also another packages can do it, like psych (REVELLE; REVELLE, 2015), vcd, (FRIENDLY, 2016), etc.

The results are present in two parts: first, the global results for sample comparing the palatal and alveolar stop realizations of /t,d/ in regressive and progressive contexts, followed by the distribution considering sex/gender stratification.

While in the rest of the country the palatal realization in regressive context is a sociolinguistic indicator, previous production studies suggest that in Sergipe this variant behaves as a positive stereotype, because it is well-evaluated and conforms to patterns in the rest of the country, and it is locally associated to “be outside of Sergipe”. The rate of application of palatal realization of /t,d/ is 12% (SOUZA NETO, 2008; SOUZA 2016) and it is in increasing process of change in the community, leaded by women, more educated, youngest and urban people. But, in progressive contexts, the palatal realization of /t,d/ is considered ‘ugly’, typically from “Nordestinos” and the countryside. Previous production studies pointed a rate of 12%, in decreasing process of change in the community leaded by men, less educated, oldest and non-urban people. In the same mood, these cues suggest that the palatal realization of /t,d/ in regressive context is recognized as a negative stereotype in the community. These cues affect how the variation is processed in terms of subjective reaction which reflect the social identity in the community, and it is expected that the pattern of judgement reflect this.

Table 1 presents the results of progressive contexts, with 11 subjects, 36 raters and 396 observations each one. The mean of percent agreement for the alveolar stop realization is 61,3% (sd = 0,06) and for palatal realization is 65,5% (sd = 0,07) in all features rated, and the range between the global means of the innovative (palatal) and conservative (alveolar stop) realization of /t,d/ in this context is 4,1. It is *quasi*-chance; and the Kappa coefficient shows that in this pairwise, the aesthetical feature “beautiful” and the rhythm feature “quick” present minimum of agreement based on the guidelines from Landis and Koch (1977), for both variables. Elsewhere contexts, there is no agreement. Kappa results reinforce that the variation in regressive contexts is not stigmatized, since the judgements are no agreement, nearby chance. It means, it is possible that the community doesn’t care if /t,d/ are palatals or alveolar stops in regressive contexts. Furthermore, all Kappa coefficients for regressive are statistically significantly different from zero (p < 0,05).

**Table 1:** Percent agreement and Kappa coefficient by regressive contexts ratings.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | alveolar stop /t,d/ | | regressive palatal /t,d/ | |
| Aesthetical features | pleasant | 0,199 | 62,3% | 0,173 | 68,2% |
| beautiful | 0,158 | 52,5% | 0,134 | 61,9% |
| clear | 0,314 | 62,6% | 0,212 | 73,0% |
| Rhythm features | quick | 0,200 | 65,5% | 0,362 | 62,1% |
| unsinging | 0,015 | 55,8% | 0,127 | 52,8% |
| Regional features | region of residence | 0,126 | 68,9% | 0,051 | 74,5% |
|  | ***mean*** |  | 61,3% |  | 65,4% |

The range in global mean for progressive contexts (with 7 subjects, 36 raters and 352 observation for alveolar stops, and 5 subjects, 36 raters and 180 observations for palatal) is 18,1 (table 2), with 63,9% (sd = 0,02) for alveolar stop realizations and 45,3% (sd= 0,14) for palatal realizations. This result signs that the negative stereotype is perceived by the judges.

**Table 2:** Percent agreement and Kappa coefficient by progressive contexts ratings.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | alveolar stop | | progressive palatal | |
| Aesthetical features | pleasant | 0,228 | 65,5% | 0,066 | 39,4% |
| beautiful | 0,225 | 60,7% | 0,059 | 32,8% |
| clear | 0,229 | 67,1% | 0,306 | 43,3% |
| Rhythm features | quick | 0,082 | 60,7% | 0,356 | 72,2% |
| unsinging | 0,058 | 64,3% | 0,051 | 51,1% |
| Regional features | region of residence | 0,076 | 65,1% | *0,025* | 32,8% |
|  | ***mean*** |  | 63,9% |  | 45,3% |

(*italics* means p > 0.05)

Kappa coefficient shows that in this pairwise, the aesthetical feature “beautiful” and the rhythm feature “quick” present minimum of agreement based on the guidelines from Landis and Koch (1977), for palatal realization; for alveolar stop realization, also the rhythm feature “quick” presents minimum agreement, as well the aesthetical features “pleasant” and “beautiful”.

In all other features in both contexts (progressive and regressive) the Kappa coefficient points to the 0, and this indicates agreement being no better than chance. Negative values indicate worse than chance agreement. It suggests, toward a sociolinguistic approach, that there is no sensibility in the community for this variation process.

Besides percent agreement > 50%, the k =0 or nearby 0 can indicate that the positive or negative values of the variables is not well detected by the raters (and, by extension in the community), or that the raters in the sample are heterogeneous.

Sex/gender is a sociodemographic feature stratified in the sample; next tables present the results separate by linguistic context (regressive and progressive) and by sex/gender.

**Table 3**: Percent agreement and Kappa coefficient by regressive contexts ratings and sex/gender raters.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | alveolar stop | | | | regressive palatal | | | |
|  |  | men | | women | | men | | women | |
| Aesthetical features | pleasant | 0,086 | 58,1% | 0,307 | 61,8% | 0,120 | 61,1% | 0,209 | 74,2% |
| beautiful | 0,106 | 50,5% | 0,212 | 54,5% | 0,071 | 55,6% | 0,202 | 68,2% |
| clear | 0,283 | 64,2% | 0,314 | 61,1% | 0,189 | 71,7% | 0,237 | 74,2% |
| Rhythm features | quick | 0,176 | 66,2% | 0,219 | 65,5% | 0,392 | 67,7% | 0,316 | 56,6% |
| unsinging | *0,003* | 63,6% | *0,008* | 48,0% | 0,149 | 61,1% | 0,148 | 44,4% |
| Regional features | region of residence | 0,126 | 66,7% | 0,142 | 52,5% | *0,040* | 77,8% | 0,050 | 71,2% |
|  | mean |  |  |  |  |  |  |  |  |

(*italics* means p > 0.05)

**Table 4: Percent agreement and Kappa coefficient by regressive contexts ratings and sex/gender raters.**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | |  | alveolar stop | | | | | progressive palatal | | | |
|  |  | | | men | | women | | men | | women | |
| Aesthetical features | pleasant | | | 0,202 | 63,5% | 0,224 | 67,5% | *0,023* | 43,3% | 0,081 | 35,6% |
| beautiful | | | 0,186 | 57,9% | 0,239 | 63,5% | *-0,035* | 32,2% | 0,147 | 33,3% |
| clear | | | 0,332 | 68,3% | 0,250 | 65,9% | 0,201 | 43,8% | 0,409 | 38,9% |
| Rhythm features | quick | | | *-0,0002* | 56,3% | 0,211 | 65,1% | 0,309 | 74,4% | 0,374 | 70,0% |
| unsinging | | | *0,042* | 69,8% | 0,067 | 58,7% | *0,016* | 48,9% | *0,039* | 53,3% |
| Regional features | region of residence | | | *0,058* | 69,0% | 0,070 | 61,1% | *0,060* | 37,8% | *-0,032* | 27,8% |
|  | ,ean | | |  | 61,5% |  | 57,2% |  | 65,8% |  | 64,8% |

(*italics* means p > 0.05)

Table 3 and 4 presents the same results from table 1 and 2, respectively, outspread by sex/gender. In the regressive context, there are 11 subjects and 18 raters for each one realization; in progressive context, there are 5 subjects and 18 raters for palatal realization and 7 subjects and 18 raters for alveolar stop realizations; the number of raters is half of the previous.

In regressive context, the mean range between alveolar stop and palatal realization for women judges (7,6) is wider than that for men judges (4,3). Women judges present more features with minimal agreement (all the aesthetical features and the rhythm feature “quick”, for both realizations, alveolar stop and palatal). Men judges present minimal agreement only in the aesthetical feature “clear”, for alveolar stop realization, and in the rhythm feature “quick”.

In progressive context, again, the mean range for women judges (20,5) is wider than that for men judges (17,4).

In this context, the realizations have a strong difference about the social value: while the alveolar stop is the non-marked realization, neutral, the palatal realization is a negative stereotype. And it is in this realization that the Kappa coefficients for men and women are convergent in the limit between the minimal and weak agreement regarding the aesthetical feature “clear” and the rhythm feature “quick”. For the alveolar stop realization in progressive context, women raters present the same pattern shown in regressive context; the same cannot said about the men raters.

When the raters are separated, negative Kappa coefficients appear (even it isn’t significate; italic values in tables are p > 0,05). According Landis and Koch (1977), a small negative Kappa coefficient should be interpreted as indicating no agreement.

No agreement seems the conclusion about these findings. The range of agreement inter-raters (<0 – 0,40) reaches the begin of weak agreement according Landis and Koch (1977) scale. The range between the contexts vary and allow us make some hypothesis about the relationship between the size and the homogeneity of sample: less raters = more range? or same group = more agreement? These questions must be explained concerning the sample size effect: how many raters must be a perception study in Sociolinguistics? Following LeBreton; Senter (2008), to calculate inter-rater agreement or reliability a sample with 10 judges is enough. But for sociolinguistic approaches it is not clearly yet (FREITAG, 2017).

Direction and strength of the dis(agree) can show directions of variation? If the linguistic change is in the community, the pattern of the judges should be agreed. But not only the times it happens.

Kappa-statistic in perception studies is a tool to observe whether there is non-homogeneous agreement between different groups of raters. It is especially important before a Principle Components Analysis (PCA) application. PCA is a data-reduction method which aims to maximize the amount of variance in the original data in a data set with fewer variables, each of which is a linear combination of the original variables. If all the raters strongly agree with each other, their ratings should all load strongly onto a single factor. So, despite the minimum agreement expected, Kappa test can be used combined with another statistic tools in order to guarantee the reliability of ratings.

**Conclusion**

Inter-rater agreement reflects the degree that different raters are interchangeable. And inter-rater reliability is measuring the relative consistency among raters. Kappa-statistic associated at another statistic tools can contribute to reach inter-rater reliability and agreement in sociolinguistic approaches to perception.

**Acknowledgments**

To Julian Tejada and Livia Oushiro for encouragement and early suggestions in this paper. To FAPITEC (PRAPEC 3/2017) for grants to attend to 46 NWAV Conference, in 2017.

**REFERENCES**

AMAN, Saima; SZPAKOWICZ, Stan. Identifying expressions of emotion in text. In: **International Conference on Text, Speech and Dialogue**. Springer, Berlin, Heidelberg, 2007. p. 196-205.

BATTISTI, Elisa et al. Palatalização das oclusivas alveolares e a rede social dos informantes. **Revista Virtual de Estudos da Linguagem**, v. 5, n. 9, p. 1-29, 2007.

BATTISTI, Elisa; DORNELLES FILHO, Adalberto Ayjara. Análise em tempo real da palatalização de/t/e/d/no português falado em uma comunidade ítalo-brasileira. **Revista da ABRALIN**, v. 14, n. 1, 2015.

BISOL, Leda. Palatalization and its variable restriction. **International Journal of the Sociology of Language**, v. 89, n. 1, p. 107-124, 1991.

CARDOSO, Denise Porto. **Atitudes linguísticas e avaliações subjetivas de alguns dialetos brasileiros**. São Paulo: Blucher, 2015.

CARLETTA, Jean. Assessing agreement on classification tasks: the kappa statistic. **Computational linguistics**, v. 22, n. 2, p. 249-254, 1996.

CARVALHO, Ana Maria. I speak like the guys on TV: palatalization and the urbanization of Uruguayan Portuguese. **Language variation and change**, v. 16, n. 2, p. 127-151, 2004.

CASTAÑEDA, Rosa-Maria. The Sociolinguistic Evolution of a Sound Change. **Journal of Portuguese Linguistics**, v. 15, 2016.

CHOUDHARY, Pankaj K.; NAGARAJA, Haikady N. **Measuring Agreement: Models, Methods, and Applications**. John Wiley & Sons, 2017.

COHEN, Jacob. A coefficient of agreement for nominal scales. **Educational and psychological measurement**, v. 20, n. 1, p. 37-46, 1960.

DAILEY, René M.; GILES, Howard; JANSMA, Laura L. Language attitudes in an Anglo-Hispanic context: The role of the linguistic landscape. **Language & Communication**, v. 25, n. 1, p. 27-38, 2005.

EDDINGTON, David. A comparison of two tools for analyzing linguistic data: logistic regression and decision trees. **Italian Journal of Linguistics**, v. 22, n. 2, p. 265-286, 2010.

EKMAN, Paul et al. Universals and cultural differences in the judgments of facial expressions of emotion. **Journal of personality and social psychology**, v. 53, n. 4, p. 712, 1987

EUGENIO, Barbara Di; GLASS, Michael. The kappa statistic: A second look. **Computational linguistics**, v. 30, n. 1, p. 95-101, 2004.

FLEISS, Joseph L. Measuring nominal scale agreement among many raters. **Psychological bulletin**, v. 76, n. 5, p. 378, 1971.

FLEISS, Joseph L.; COHEN, Jacob. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. **Educational and psychological measurement**, v. 33, n. 3, p. 613-619, 1973.

FREITAG, Raquel Meister Ko; SOUZA, Gládisson Garcia Aragão. O caráter gradiente vs. discreto na palatalização de oclusivas em Sergipe. **Tabuleiro de Letras**, v. 10, n. 2, p. 78-89, 2017.

FREITAG, Raquel Meister Ko. Banco de dados falares sergipanos. **Working Papers em Linguística**, v. 14, n. 2, p. 156-164, 2013.

FREITAG, Raquel Meister Ko. Amostras sociolinguísticas: probabilísticas ou por conveniência?. **Revista de Estudos da Linguagem**, v. 26, n. 2, p. 667-686, 2017.

FREITAG, Raquel Meister Ko. Socio-stylistic aspects of linguistic variation: schooling and monitoring effects. **Acta Scientiarum. Language and Culture**, v. 37, n. 2, p. 127, 2015b.

FREITAG, Raquel Meister Ko.; SANTOS, Adelmileise de Oliveira. Percepção e atitudes linguísticas em relação às africadas pós-alveolares em Sergipe. In: LOPES, Norma da Silva; ARAÚJO, Silvana Silva de Farias; FREITAG, Raquel Meister Ko. **A Fala Nordestina: entre a sociolinguística e a dialetologia**. São Paulo: Blucher, 2016, p. 109-122.

FRIENDLY, Michael. Working with categorical data with R and the vcd and vcdExtra packages. **The Comprehensive R Archive Network**, 2013.

LABOV, William. **Sociolinguistic patterns**. Pennsylvania: University of Pennsylvania Press, 1972.

LABOV, William. **The social stratification of English in New York city**. Cambridge: Cambridge University Press, 2006.

LADEGAARD, Hans J. Language attitudes and sociolinguistic behaviour: Exploring attitude‐behaviour relations in language. **Journal of Sociolinguistics**, v. 4, n. 2, p. 214-233, 2000.

LANDIS, J. Richard; KOCH, Gary G. The measurement of observer agreement for categorical data. **biometrics**, p. 159-174, 1977.

LEBRETON, James M.; SENTER, Jenell L. Answers to 20 questions about interrater reliability and interrater agreement. **Organizational research methods**, v. 11, n. 4, p. 815-852, 2008.

MCHUGH, Mary L. Interrater reliability: the kappa statistic. **Biochemia medica: Biochemia medica**, v. 22, n. 3, p. 276-282, 2012.

MOTA, Jacyra Andrade. Como fala o nordestino: a variação fônica nos dados do Projeto Atlas Lingüístico do Brasil. In: LIMA-HERNANDES, MC et al. (Org.). **A língua portuguesa no mundo.** São Paulo: FFLCH-USP, 2008.

OUSHIRO, Livia. **Salience and covariation in second dialect acquisition: Northeastern migrants in São Paulo**. Paper presented at New Ways of Analyzing Variation 46, 2017.

PAOLILLO, John C. **Analyzing linguistic variation: Statistical models and methods.** Chicago: Center for the Study of Language and Information, 2002.

PEIRCE, Jonathan W. PsychoPy—psychophysics software in Python. **Journal of neuroscience methods**, v. 162, n. 1-2, p. 8-13, 2007.

POSNER, Karen L. et al. Measuring interrater reliability among multiple raters: an example of methods for nominal data. **Statistics in medicine**, v. 9, n. 9, p. 1103-1115, 1990.

R CORE TEAM. **R: A language and environment for statistical computing.** 2018.

REVELLE, William; REVELLE, Maintainer William. Package ‘psych’. **The Comprehensive R Archive Network**, 2015.

SOUZA NETO, Antônio Félix de. **Realizações dos fonemas/t/e/d/em Aracaju Sergipe**. 2008. Dissertação (Mestrado em Letras). Universidade Federal de Alagoas, 2008.

SOUZA, Gládisson Garcia Aragão et al. **Palatalização de oclusivas alveolares em Sergipe.** 2016. Dissertação (Mestrado em Letras). Universidade Federal de Sergipe, 2016.

TAGLIAMONTE, Sali A.; BAAYEN, R. Harald. Models, forests, and trees of York English: Was/were variation as a case study for statistical practice. **Language variation and change**, v. 24, n. 2, p. 135-178, 2012.

TINSLEY, Howard; WEISS, David. **Handbook of applied multivariate statistics and mathematical modeling**. Academic Press, 2000.

TINSLEY, Howard; WEISS, David. Interrater reliability and agreement of subjective judgments. **Journal of Counseling Psychology**, v. 22, n. 4, p. 358, 1975.

1. Universidade Federal de Sergipe, Centro de Educação e Ciências Humanas, São Cristóvão, Sergipe, Brasil. Professora do Departamento de Letras Vernáculas. E-mail: [rkofreitag@uol.com.br](mailto:rkofreitag@uol.com.br) [↑](#footnote-ref-1)
2. A previous version of this paper was present at 46 NWAV Conference (Madison, WI, United States), in 2017, with grants from FAPITEC (PRAPEC 3/2017) and at Sociolinguistic meeting in ENANPOLL (2018). [↑](#footnote-ref-2)