Kappa statistic for judgment agreement in Sociolinguistics

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 through 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 explanatory power of perceptual studies in Sociolinguistic.

KEYWORDS: Perception study; statistic tools; inter-rater agreement; palatalization.

*Estatística Kappa para concordância de julgamento em Sociolinguística*

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**

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

Considering the call for papers of Relin about new ways of analyzing perception in Sociolinguistics, this paper provides a proposal for a new look into the quantitative approach of data collected in perception studies in Sociolinguistics, as happened with the first studies of production in Sociolinguistics. Methodological discussions about how it is possible to get more statistical significance to quantitative approaches in Sociolinguistics have been held since the first studies, for example, the quantitative approach to variation of /R/ in department stores in New York city, which was initially explored only in percentage (LABOV, 1972, 2006), and afterward in logistic regression with Varbrul (PAULILLO, 2002), or logistic regression and decision tree methods (EDDINGTON, 2010; TAGLIAMONTE; BAAYEN, 2012). Each one of these approaches with the same dataset contributed to design the pattern of analysis currently adopted for data from production studies in Sociolinguistics. 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, as in the pioneering production studies.

This text aims at: 1) exploring the concepts of reliability and agreement in judgment tests (as in subjective reaction test, for example), considering the inherent inter-raters and intra-raters variance, 2) presenting a mathematical model to quantify the variability in judgment tests, the Kappa statistic, and 3) reanalyzing the previous dataset using Kappa statistics, in order to improve the perceptual approaches in Sociolinguistics.

**Quantitative approach to perceptual studies in Sociolinguistics**

Perception studies in Sociolinguistic try to answer questions about the role of speakers, which presupposes the inference of a certain direction toward change from 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 sum of answers is not necessarily consensual: if all the speakers answered “quick”, or “ugly”, there is a consensus, and it is regarded as a complete change. In incoming change processes, the variability in answers is expected and desirable; besides, identifying the degree of agreement in the answers can help to interpret it as a direction of the linguistic change. It enables the identification of the sex/gender, dialectal, educational, or other social or cultural bias driving the variant choices. For example, women (either educated, or urban) can attribute more positive values to one variant than men (either non-educated, or rural people).

One technique to account for it is the percentage of judgments as a measure of agreement. It is apparently easy to calculate it and to interpret it because it consists in the computing of occurrences to each feature rated by judges. However, the computing of percentage does not account for the agreement that is consequence from random variability or chance. Fieldwork that studies the effects of pattern of judgment, as psychology, education, medicine, etc. has adopted Kappa statistics to consider the effect of random variability. In this paper, there is a proposal to broaden this scope also to cover the perception studies in Sociolinguistics. The concepts of reliability and agreement in psychometrics are explained in order to detail Kappa statistic and its applications in perceptual approaches in Sociolinguistics.

***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 the relative consistency of a measuring, whether a test, a scale, or a pattern is consistent. It is different from agreement, which refers to the convergence of the results.

This sum considers the intra-rater reliability (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 of ratings, or the degree of agreement between the choices made by two or more independent judges) (LEBRETON; SENTER, 2008). It means, whether judges rank other targets in a manner that is relatively consistent with themselves and with other judges, not only in scores, but also in 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 any person or entity in action, like anesthesiologists, psychiatrists, nurses, etc.) and the subject/target of the rating can be people, things, processes, outcomes, data (time since the 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). The Kappa statistics is then 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 subject 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-weighted Kappa), as seen above. The coefficient expansion is labeled Kappa-like statistics or Kappa statistics (POSNER et al, 1990).

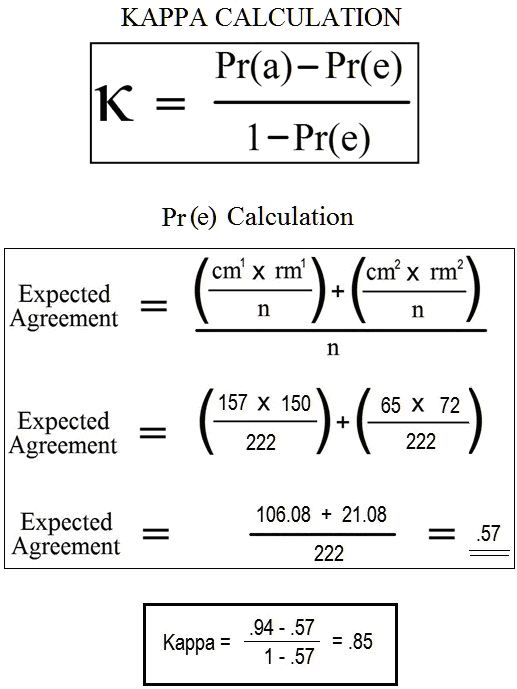
While Cohen’s Kappa is an agreement measure based on the pattern of two raters in a categorical scale for the subjects, Fleiss’s Kappa is applied among more than two fixed or random raters. Expanding the scope, weighted Kappa (FLEISS; COHEN, 1973) is an agreement measure to ordinal data (CHOUDHARY; NAGAJARA, 2017). Thus, the choice among specific Kappa test depends on the amount of 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 for Kappa to be negative, which means that the two raters agreed less than it would be expected just by chance (and it is particularly important in Sociolinguistics, 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) 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 are 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 that all the raters assess the same observations; 3) The raters are independent, it means that one rater's judgment does not affect the other raters’ judgment.

Kappa statistic has several applications in research fields where judgment tasks are required. It is also and used 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), for example, 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, which includes attitudes, judgments, and emotions. The procedure consisted in comparing the results of one judge, the specialist who established the gold standard, against the other three, lay participants, and calculating the mean of agreement with gold standard 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) also adopted Kappa statistic in a judgment test which consisted of the selection of only one emotion term for each expression, in a set of three expressions per each emotion (180 subjects) by 10 raters for each one of the 10 cultures.

Carletta (1996, p. 253) points to the interest of computational linguistics and cognitive science in subjective judgments. She claims the Kappa statistic is 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 disagreement 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 impressionistically cut off. 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 a palatalization process that produces a gradience of sounds, for example between /t/ and /tʃ/ in Brazilian Portuguese (XXXXX, YYYYYY 2016).

Suppose a task in which two transcribers were given a set of 40 sounds, and they must rate each one of them in /t/ or /tʃ/. There is no gold standard which 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/. Then, the result based on these rates can provide a gold standard for validating next transcriptions. To constitute a dataset, the judge’s 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 judgment for all the 40 sounds, as seen in the column “agreement” in the spreadsheet 1.

Both transcribers agree that 25 sounds from the dataset are alveolar stops and 15 are palatals. Based on these transcribers’ judgments in the dataset, the rate of palatalization is 63%, and 38% for alveolar stops. In addition to that, both transcribers totally agree (100% of agreement). This is an ideal and unlikely scenario, as the next one, from spreadsheet 2, where the distribution of the rates is equal: transcriber 1 and 2 randomly rating 50/50 of the sounds in t or tʃ, and the rate of palatalization is the same as 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, the transcriber in disagreement must be identified. It is not possible to identify them only based with the percentage of agreement. To do it, one more rater must be added to compare their mean of agreement with the other ones, then deciding who the transcriber is in disagreement. To correct this mistake, the Kappa statistic is required. The first step is transforming the data in the spreadsheets in an *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 to the calculations for the expected agreement Pr(e), and observed agreement Pr(a). The observed agreement is calculated by the sum of the frequencies in the main diagonal cells (cells a and d) divided by n, as in (2). When multiplying by 100, the percentage agreement is found.

(2)

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

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 to be /t/ or /tʃ/ is calculated by multiplying the marginal totals corresponding to each cell on the main diagonal and dividing by n. The proportion of expected agreement is then 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

In the third step, these results are applied 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 in figure 3 points to the perfect agreement, and spreadsheet 2 in figure 3 points to the randomly agreement.

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



As it seems obvious, the perfect agreement has the highest Kappa coefficient and the randomly agreement is 0.0, in the level of non-agreement, according Landis and Koch’s (1977) guidelines presented in figure 2. However, the results are not always as 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, the percentage of agreement is 60% and κ = 0.20, suggesting no agreement; in D, the percentage of agreement is 40% and κ = -0.20. McHugh (2012) warns that Kappa coefficient 0 above 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 the other one is lay, for example). A negative Kappa signs that agreement is worse than expected, or disagreement. On the other hand, 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. Nevertheless, it is particularly interesting for perception studies in Sociolinguistics (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 between transcribers.



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

**Reanalyzing a perceptual study with Kappa statistic**

Seeking to reach the goals of this paper, a new analysis of the dataset from another study was carried out. The research question concerns the perception of how /t/ and /d/ sound in Sergipe: palatal or alveolar stop. Palatalization of /t/ and /d/ followed by the vowel /i/, like in “tia” (“aunt”) and “dia” (“day”), is called regressive palatalization. There is another process in Brazilian Portuguese, when the /t/ and /d/ is preceded by the glide /y/, it is called progressive palatalization, like in “oito” (“eight”) and “peito” (“chest”).

In regressive contexts, the standard of urban dialects almost all over Brazil is the palatal /t/ and /d/ realization. In some places, the alveolar stop realization is more productive, which is related to certain dialectal scopes and certain indexical fields. For instance, 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). The same happens in the dialectal scope of southeastern Brazil, where alveolar stop realization is related to being “caipira” and “nordestino”, people who come from the northeastern region (OUSHIRO, 2016). In both regions, the value associated to the alveolar stop realization is negative. On the other hand, progressive palatalization is less recurrent, and highly stigmatized, and more restricted to certain dialectal areas like the countryside of Sergipe (MOTA, 2008, XXXXXXX, 2015).

The results of production studies in Sergipe suggest a change in progress, but these studies do not inform the social forces driving this process nor how the innovative variant is evaluated by community.

XXXXX & YYYYYY (2016) presented results of undergraduates’ perception about the variation in /t,d/ realization in Brazilian Portuguese, in a dialectal area where the change is happening in Sergipe, Brazil. This previous analysis was strictly exploratory and considered only the percentage of responses, without any 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 YYYYYY YYYYYY YY YYYYYY database (XXXXXXX, 2013), in conditions near to minimal pairs of isolated words (LADEGAARD, 2000; DAILEY; GILES; JANSMA, 2005), without manipulation.

The task consisted of the judges hearing the stimuli to respond a sequence of questions regarding aesthetical, rhythm, and regional features about the way of speaking, following 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 from YYYYYY YYYYYY YY YYYYYY who were born and live in the capital, and are surrounded by an urban region, stratified equally in male and female. The rating task was run in PsychoPy v.1.82.01 (PEIRCE, 2007), in the XXXXXXXX XXXXXXX XX XXXXXXXX X room, at YYYYYY YYYYYY YY YYYYYY.

As explained, the sample is relatively homogeneous, composed of undergraduates (same educational level and presumably the same age group). Therefore, only the raters’ sex/gender was controlled. Intending to expand the discussion beyond the percentage of responses, measuring the agreement among the raters and the convergence of ratings in the sample, the following questions guide this new analysis: Does the percentage of ratings match the expected agreement?; Do the observed agreement and the percentage of ratings converge?; Does the sex/gender bias affect the perception results?

There are 36 raters, each one independently rated all stimuli once, 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 sum of observed and expected frequencies to calculate the Fleiss’s Kappa. It can be made using a spreadsheet software, like ®Microsoft Excel, or a statistical software, like ®SPSS or R (R CORE TEAM, 2018). Finally, Fleiss’s Kappa is calculated with irr package, kappam.fleiss function (GAMER, 2016), but other packages can do it too, like psych (REVELLE; REVELLE, 2015), vcd, (FRIENDLY, 2016), etc.

The results are presented in two parts: first, the global results for the sample, comparing the palatal and alveolar stop realizations of /t,d/ in regressive and progressive contexts, and second 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, led by women, more educated, younger, and urban people. However, in progressive contexts, the palatal realization of /t,d/ is considered ‘ugly’, typically from “nordestinos” and people from the countryside. Previous production studies pointed to a rate of 12%, in decreasing process of change in the community led by men, less educated, older, and non-urban people. These cues suggest that the palatal realization of /t,d/ in regressive context is recognized as a negative stereotype in the community. They also affect how the variation is processed in terms of subjective reaction which reflects the social identity in the community. Therefore, the pattern of judgment is expected to reflect this.

Table 1 presents the results of progressive contexts, with 11 subjects, 36 raters, and 396 observations each. The mean of the percentage of agreement for the alveolar stop realization is 61.3% (sd = 0.06) and for the 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. 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. In other contexts, there is no agreement. Kappa results reinforce that the variation in regressive contexts is not stigmatized, since the judges do not agree, it is close to chance. It means that the community may not care if /t,d/ are palatals or alveolar stops in regressive contexts. It is out of awareness. Furthermore, all Kappa coefficients for regressive contexts are statistically significantly different from zero (p < 0.05).

**Table 1:** Percentage of agreement and Kappa coefficient for 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% |
| not sung | 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 639% (sd = 0.02) for alveolar stop realizations and 45.3% (sd= 0.14) for palatal realizations. This result indicates that the negative stereotype is perceived by the judges.

**Table 2:** Percentage of agreement and Kappa coefficient for progressive context 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% |
| not sung | 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 agreement for palatal realization, based on the guidelines from Landis and Koch (1977). For alveolar stop realization, the rhythm feature “quick” also presents minimum agreement, as well as the aesthetical features “pleasant” and “beautiful”.

In all other features in both contexts (progressive and regressive), the Kappa coefficient points to 0. This indicates that the agreement is not better than chance. Negative values indicate that agreement is worse than chance (0.5). Toward a sociolinguistic approach, it suggests that there is no sensibility in the community for this variation process.

Besides, the percentage agreement > 50%, the k =0 or near 0 may indicate that the positive or negative values of the variables are not well detected by the raters (what may be extended to the community), or that the raters in the sample are heterogeneous.

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

**Table 3**: Percentage of agreement and Kappa coefficient for regressive context 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% |
| not sung | *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% | 0050 | 71.2% |
|  | Mean |  |  |  |  |  |  |  |  |

(*italics* means p > 0.05)

**Table 4: Percentage of agreement and Kappa coefficient for regressive context 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% |
|  | mean | |  | 61.5% |  | 57.2% |  | 65.8% |  | 64.8% |

(*italics* means p > 0.05)

Tables 3 and 4 present the same results as tables 1 and 2, respectively, outspread by sex/gender. In the regressive context, there are 11 subjects and 18 raters for each 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 sample (tables 1 and 2).

In regressive context, the mean of the range between alveolar stop and palatal realization for women judges (7.6) is higher 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). Men judges present minimal agreement only in the aesthetical feature “clear”, for alveolar stop realization, and in the rhythm feature “quick”, for palatal realization.

In progressive context, again, the mean range for women judges (20.5) is higher 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. 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 (despite not being significant; italic values in tables are p > 0.05). According to Landis and Koch (1977), a small negative Kappa coefficient should be interpreted as indicating no agreement.

No agreement seems to be the conclusion about these findings. The range of agreement inter-raters (<0 – 0.40) reaches the beginning of weak agreement according to Landis and Koch’s (1977) scale. The range between the contexts varies and allows us to make some hypothesis about the relation between the size and the homogeneity of the sample: less raters = more range? or same group = more agreement? These questions must be explained concerning the sample size effect: how many raters there must be in a perception study in Sociolinguistics? Following LeBreton; Senter (2008), to calculate inter-rater agreement or reliability, a sample with 10 judges is enough. Nevertheless, for sociolinguistic approaches it is not clearly yet (XXXXXX, 2017).

May the direction and strength of the (dis)agreement show directions of variation? If the linguistic change is in the community, the pattern of the judges should agree, but that does not always happen.

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

**Conclusion**

Inter-rater agreement reflects the degree that different raters are interchangeable. Besides, inter-rater reliability measures the relative consistency among raters. Kappa statistic associated with other statistic tools can contribute to reach inter-rater reliability and agreement in sociolinguistic approaches to perception.

**Acknowledgments**

XXX

**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.

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.

XXXXX

YYYYY