Quantitative models for the study of linguistic variation

ANTHONY NARO
Universidade Federal do Rio de Janeiro

The Portuguese language has an inherent series of restrictions, which make certain strings of words, such as *a casa* [the house] perfectly normal, while other strings, such as *casa a* [house the] are considered anomalous. [1] Such categorical restrictions are so strong that the non-occurring string is virtually impossible. In other cases, the language possesses two or more variant forms that can be used by the speaker without causing any noticeable change in the message transmitted. This is especially true in phonology, where we find variants such as *pe**i**xe/pexe* [fish] (with or without a diphthong), *home**m**/home* [man] (with or without nasalization), *m**e**nino/m**i**nino* [boy] (with a mid or high vowel in the first syllable). In morphology, there are similar examples, such as double participles of the type *pego/peg**ad**o* [caught] or agreement within the noun phrase in informal speech (*uma**s** casinha**s** bonitinha**s** / uma**s** casinha bonitinha* [some pretty little houses]).

In reference to syntax, the variants sometimes carry a slight change in the message transmitted, but nonetheless still represent options for the speaker, requiring such choices as the active or the passive (*alguém lavou os pratos* [someone washed the dishes] / *os pratos foram lavados* [the dishes were washed]), the verb-subject or the subject-verb order (*apareceram três homens* [there appeared three men] / *três homens apareceram* [three men appeared]), or the use or omission of a subject pronoun (*ele chegou* [he arrived]/ *chegou* [arrived]). Even at the level of discourse, the speaker must confront a series of choices, which can often be quite difficult, such as the appropriate way to address the hearer (*tu*, *você*, *o senhor*, or simply the omission of any particular pronoun, all corresponding to *you* in English).

The most basic assumption in the study of linguistic variation is that linguistic heterogeneity, like linguistic homogeneity is not random, but regular or rule-governed. In other words, just as there are CATEGORICAL RULES that require the speaker to consistently use certain forms (*a casa* [the house]) instead of others (*casa a* [house the]). There are also VARIABLE RULES that favor to a greater or lesser degree the use of one form over another in each environment in accordance with well defined relative weights. This means that language variants compete with each other. Accordingly, we postulate that even though one of the variants may sometimes occur in place of the others, it should be possible to identify a series of independent categories that influence the outcome of the choice of the dependent variable. These categories can be found both within the linguistic system itself, as well as outside of it. In the first case we shall be dealing with structural factors, such as the relatively less frequent use of the diphthong *ei* when the following segment is palatal (e.g. *pexe [peši]* more frequent than *peixe [peiši]* [fish]) in comparison to when it is alveo-dental (e.g. *peito* more frequent than *peto* [breast]). In the second case, social factors also come into play. For example, people with more formal education use agreement more frequently than illiterate or semi-literate speakers.

The most important social factors seem to be age, sex, socio-economic level, and level of education. Other relevant social factors in certain cases of variable linguistic usage are the speaker's position in the labor market and his degree of contact with the mass media (television, newspapers, etc.). However, each phenomenon must be studied in the context of its own particular social embedding. For example, in the case of pronouns corresponding to `you', relevant factors are differences in the social status and age of the speaker and the interlocutor, as well as the degree of formality characteristic of the speaker in general.

In the case of structural factors, each phenomenon must be considered in light of its relationship to the relevant wider structural context. For subject/verb agreement, two of the most important structural categories are:

1) SALIENCE of the singular/plural opposition: When the phonic difference between the singular and plural forms is small, agreement

marking is not favored; when the phonic difference is greater, agreement marking is favored. Thus, in the opposition *eles bebe / eles bebem* [they drink] (opposition -[i]/-[_]) the first form, without a surface plural marking, is more favored than in *eles bebeu / eles beberam* (opposition [éu]/[éru]) [they drank].

2) POSITION of the subject with respect to the verb: When the subject comes immediately before the verb, agreement is favored; but when the subject comes after or is more distant from the verb, the opposite is true. Thus, the singular surface form is less favored in *os embaixadores chegaram* [the ambassadors arrived] than in *chegou em território nacional, após o término das negociações sobre a dívida externa em Nova Iorque, os embaixadores brasileiros* [the Brazilian ambassadors, after the conclusion of the negotiations on the external debt in New York, arrived in Brazilian territory].

The main problem in any theory of linguistic variation is evaluating how much each independent category contributes to the use of one dependent competing variant form over another. In linguistic data, such as real speech, these categories are always found in conjunction with others because in practice, the use of a linguistic variable is always the result of the simultaneous effect of several factors. In the case of subject/verb agreement, all occurrences of a third person plural verb form, whether formally marked for plural or not, will belong to one of the morphological categories and one of the positional categories. In real data, it will never be possible to check the effect of the morphological category in isolation from the simultaneous effect of the positional category. Both will always be obligatorily present, since any verb that can occur in the third-person plural must have a subject of some kind, no matter whether it is preposed, postposed, deleted, or separated from the verb.

In view of this, it is, at least in principle, impossible to measure the influence of a given linguistic category directly in real speech data without at the same time, measuring the effect of other categories that are also present. In other words, the role of variable rule theory in the study of linguistic variation is to isolate and measure the effect of a given factor (e.g. the fact that the subject is immediately preposed to the verb) even though the factor can never be isolated in the data (e.g. the verb to which the subject is immediately preposed will always simultaneously display some morphological category). The property `immediately preposed subject' is an analytical abstraction that is realized only in conjunction with other properties. Similarly, in the study of the effect of a speaker's gender, it is highly desirable to separate such an effect from that of other factors, such as age and educational level. However, speakers are an indivisible complex of all of these features, and obviously no speaker can be of a certain gender without at the same time being of a certain age, and having a certain educational level. As a result, it goes without saying that social effects cannot be separated from structural ones because our data always consist of someone saying something.

Before proceeding any further, we must stop and ask ourselves if a solution for this problem can be found. Given that there is no entity that only has the property `female', is it really possible to discover what effect being female has on one's speech? The answer undoubtedly depends on the consistency of the behavior of the category we are interested in. Suppose that in the agreement phenomena discussed above, there is no clear trend for age in the population as a whole, but we find that older men use agreement less often than younger men, while older women use agreement more often than younger women. In this case, younger women appear to be abandoning the rule in comparison to older women, while younger men appear to be using it more often in comparison to older men. Nevertheless, it is not possible to make any statement at all about the behavior of women or men independently of their age.

When the effects of two or more categories cannot be separated, they are said to INTERACT or to lack INDEPENDENCE[2] and in this case, our problem does not appear to have an easy solution. At best, instead of the categories `male', `female', `young' and `old', we could use `young female', `old female', `young male', and `old male'. In the case at hand we would be trading four simple categories for four complex ones. In general, however, for two groups with s_1 categories in the first and s_2 categories in the second we will have the product of s_1 x s_2 complex categories as opposed to the sum of $s_1 + s_2$ simple categories. The number of complex categories determined by the product can quickly become large enough to make analysis impractical when we are dealing with groups containing many factors or with more than two groups.

As a necessary preliminary to variable rule analysis, it is thus imperative to determine whether the categories under study are independent of each other. If this is the case, we can use simple categories in our analysis, but if not, we have no choice but to use complex ones.[3]

The practical importance of separating the effect of each category has its illustration in a problem that arose in the course of a recent research project. In the initial stages of their study of agreement within the noun phrase in spoken Brazilian Portuguese, Braga (1977) and Scherre (1978) believed that they had discovered a relevant morphological category, parallel to (1) above. They thought that the chances of non-occurrence of the plural morpheme were greater in those words in which the singular/plural opposition was simply *∅/s* (*casa/casaS* [house/houses]) than in those cases in which the opposition was more complex (*milhão/milhÕES* [million/millions] or *hotel/hotéIS* [hotel/hotels]). However, their hypothesis seemed doomed in view of the initial results obtained from an exploratory corpus consisting of two speakers: 49% of the simple opposition forms were found to be marked for the plural, while the forms of the more complex opposition, where more plural marking was expected, had an explicit plural morpheme in only 36% of the cases.

However, it must be taken into account that a second category had also been postulated as relevant to the rate of plural marking in noun phrases, namely, the linear position of the element within the noun phrase. Three positions were recognized, as illustrated below:

> -*Position 1 : first markable element of the NP.* Marking was present in about 98% of the cases (e.g. *A(S) minha(s) amiga(s)* [(the) my friends (f. pl.)]).
> -*Position 2: second markable element of the NP.* Marking was present in about 18% of the cases (e.g. *a(s) MINHA(S) amigas(s)* [(the) my friends (f. pl.)]).
> -*Position 3 -- third markable element of the NP.* Marking was present in only 10% of the cases (e.g. *a(s) minha(s) AMIGA(S)*

[(the) my friends (f. pl.)]).

As a result, there are at least two simultaneous factors at work in the variable use of a plural mark on an element of a noun phrase: the morphological class of the pluralizable element and its linear position within the noun phrase.

If we examine the distribution of the data for the two morphological categories, beginning with the simple opposition (*∅/-s*), the distribution of noun phrase elements with respect to the three linear positions (both with and without plural marking) is as follows:

position 1: 39%
position 2: 55%
position 3: 6%

In the case of the complex opposition, the distribution of the linear positions occupied by noun phrase elements is the following:

position 1: 0%
position 2: 91%
position 3: 9%

Since there is very little data showing position 3, positions 1 and 2 will be our main consideration. In the case of the simple opposition, there is a slightly higher concentration of tokens in position 2 (55% in comparison with 39%), but for the more complex opposition, there are no occurrences at all in position 1, and virtually all the data is thus concentrated in position 2 (0% in comparison to 91%). The reason behind this uneven distribution lies in the fact that position 1 is normally occupied by a determiner, such as an article or a demonstrative. These items all belong to the simple opposition (e.g. *a/as* [the (fem. sg./pl.)], *o/os* [the (masc. sg./pl.], *este/estes* [this/these (masc.)] etc.). In this very restricted corpus, there were no examples of the type *hotéis bonitos* [pretty hotels], with the complex opposition in position 1, because this construction is infrequent in real usage, though perfectly grammatical.

It should be emphasized that the simple opposition has a significant part (= 39%) of its occurrences in position 1, where the plural morpheme is almost always used by speakers (= 98% presence). In contrast, ALL of the tokens of the complex opposition are to be found in positions 2 and 3, where the plural mark is almost always omitted (= only 18% presence for position 2 and 10% for position 3). These distributional facts cause a spurious inflation in the overall frequency of marking in the simple opposition class. In other words, the simple opposition shows such high frequencies simply because it is concentrated in position 1, and not for any reason related to its morphological structure. Since position 1 causes problems in our interpretation of the results, this type of data has not been considered, and only data from positions 2 and 3 has been retained. The new distribution of data is as follows:

	simple opposition	*complex opposition*
position 2	90% (=55/61)	91%
position 3	10% (= 6/61)	9%

The distribution of the complex opposition does not change when position 1 data is eliminated because it had no occurrences at all in this position. However, the situation is radically different regarding the simple opposition, since in this case, 39% of the data is lost. As the table shows, the new distribution is very balanced, eliminating the problems of analysis discussed above.

Considering only the data in positions 2 and 3, the frequencies of plural marking for the two morphological categories become 36% for the complex opposition and only 17% for the simple opposition. The frequency for the complex category (36%) does not change when position 1 data is eliminated simply because this category did not occur in position 1. Nevertheless, in the case of the simple category, the frequency decreases from 49% to 17% because of the removal of the frequently marked cases of position 1, which were inflating the results. The new frequencies obtained from a non-biased sample confirm the initial hypothesis as to the greater favoring of marking by the complex opposition.

The obvious conclusion that can be drawn is that raw frequencies, though concrete and intuitively "real", can be very deceptive because they do not take into account the relationships between the categories that influence the outcome of linguistic variation. What is needed is an hypothesis that explains how the effects of the various categories in a given environment combine to produce the overall effect observed in the empirical data. With such an hypothesis, computational methods could be used to isolate the individual effects of each category.

For purposes of notation we will represent the effect of a category (in our example, morphological class and position) by means of a subscripted variable such as f_i. In each environment, one factor from each category will be present. For example, in *as casa* [the houses], the determiner *as* [the] exhibits the joint effect of the factors of `simple opposition' (= $f_1$) and `position 1' (= f_2). The overall effect in this environment will be denoted by f_t (= frequency of the use of the plural marking in the environment). The hypothesis that we are seeking should relate f_t to f_1, f_2, etc. by means of some mathematical function. In what follows, we will review several attempts that have been made by statisticians to supply a theoretical mathematical model consistent with the linguistic data, which eventually led to the logistic model, the one that is most predominantly used in quantitative sociolinguistics today.

In 1969, William Labov proposed an *additive model*, in which f_t was postulated to be the sum of the effects of the factors present in a given environment. We would thus have:

$$f_t = f_0 + f_1 + f_2 + ... + f_n$$

In this formula, n is the total number of relevant categories and f_0 is the overall frequency, or grand mean, of use of the variant under examination. In an investigation of subject/verb agreement in the speech of adult literacy students in Rio de Janeiro, Naro (1981) found a grand mean for use of the plural mark in all environments of 47.6% (=f_0). However, with the subject immediately preposed to the verb, oppositions of the simple type (*bebe / bebem*) exhibited this mark only 17.2% of the time, while oppositions of the complex type (*bebeu / beberam*) showed a formal plural mark in 76.4% of all possible cases. Under the additive model, the calculations for the values of f_i

associated with each factor resulted in the numbers given below for the categories shown:

	bebe/bebem	*bebeu/beberam*	
f_0	47.6%	47.6%	(grand mean)
f_1	-33.7%	25.5%	(salience)
f_2	3.3%	3.3%	(immediately preposed S)
TOTAL	17.2%	76.4%	

In this model, the grand mean serves as a point of reference for the effect of the factors, the value of which can be either positive or negative. Positive effects correspond to factors which tend to cause an increase in the frequency of use of the variant with respect to the grand mean, while negative effects correspond to a decrease in this frequency. In slightly more technical terms, f_0 is an initial value or so-called INPUT, and the f_i are DEVIATIONS with respect to this value. In less technical terms, we would say that the marked variant generally shows a usage frequency of 47.6% (the grand mean). However, the fact that there is a subject immediately preceding causes the usage frequency to increase by 3.3%, while the fact that the verb is of the simple category (*bebe / bebem*) decreases it by 33.7%. In contrast, when the verb belongs to the complex category (*bebeu / beberam*), there is an increase of 25.5%. These percentages (33.7%, 25.5%., and 3.3%) are calculated so as to fit the additive model as closely as possible to the empirical results (17.2% and 76.4%), as well as to those of all other environments in which these factors appear. In our example, it was possible to fit the model perfectly to the data, but this is usually not the case. Consequently, there is inevitably a margin of error to be taken into consideration in each environment.

Unfortunately, despite its great intuitive appeal, the additive model was abandoned because of technical problems that at the time, were considered insurmountable. As this model consists of a sum of positive and negative numbers, there is really no way to guarantee that the result will not be greater than 100% or less than 0%. Numbers outside this range would complicate calculations considerably, and would not, in any case, correspond to reality.

In 1974, Henrietta Cedergren and David Sankoff proposed an interpretation in which frequencies are replaced by probabilities. Whereas frequencies are empirically established and expressed as numbers within the range of 0% to 100%, probabilities are limited to the range from zero to one, and are derived from theoretical considerations. A probability of zero indicates that an event can never occur; the value 1 corresponds to an event that always occurs; and other numbers indicate events that may occur sometimes.

For example, we might toss a coin in the air 100 times and find that in 48 instances, the result was heads. We would then say that the frequency of heads in this particular experiment was 48%. To derive the probability of getting heads, we would have to take into consideration any conditions, physical or otherwise, that might influence the outcome of the toss. Possible factors that could hypothetically be taken into account are the metallic composition or density of the coin, the presence of magnetic fields or air currents, etc. Since there normally are no such conditions, we would conclude that the probability of obtaining heads is 0.5. No experimentation is required to derive this probability.

Cedergren and Sankoff suggested the classical probabilistic model of the simultaneous effect of independent factors as the basic model for the joint effect of factors, corresponding to the role attributed earlier to the additive model. If we replace *f* by *p* in our notation, the new model postulates p_t to be the PRODUCT (and not the SUM) of the factors p_i , in the same way that the probability of obtaining two heads in successive tosses of a coin is 0.25 = 0.5 x 0.5 (i. e. the product of the probabilities of having heads come up on each toss[4]). This *multiplicative applicational model* is formulated as follows:

$$p_t = p_0 \text{ x } p_1 \text{ x } p_2 \text{ x } \ldots \text{ x } p_n$$

At first sight, the multiplicative applicational model seems to be even less satisfactory than the additive model since it entails that the joint effect of two highly favorable factors be less than the effect of either one of them in isolation. This is so because the product of two numbers close to one will always be less than either one of them (e.g. for $\underline{f}_1 = 0.9$ and $\underline{f}_2 = 0.8$ the result of the product 0.9 x 0.8 is 0.72).

If *f* is the probability of occurrence of an event, such as the application of a linguistic rule, the probability of non-occurrence is *1-p*. Consequently, we can transform the multiplicative model to non-applicational by subtracting 1 from each term of the equation. The new non-applicational multiplicative model is:

$$(1-p_t) = (1-p_0) \text{ x } (1-p_1) \text{ x } (1-p_2) \text{ x } \dots \text{ x } (1-p_n).$$

This model solves the problem of the joint effect of two highly favorable factors. The relevant calculations are:

$$\begin{aligned}(1-p_t) &= (1-0.9) \text{ x } (1-0.8)\\ &= (0.1) \text{ x } (0.2)\\ &= (0.02)\end{aligned}$$

or

$$p_t = 0.98.$$

The result, 0.98, is greater than 0.9 or 0.8. However, for two highly unfavorable values, such as $p_1 = 0.1$ and $p_2 = 0.2$, the calculation would be the following:

$$\begin{aligned}(1-p_t) &= (1-0.1) \text{ x } (1-0.2)\\ &= (0.9) \text{ x } (0.8)\\ &= (0.72)\end{aligned}$$

or

$$p_t = 0.28$$

In this case, despite the fact one would expect the joint effect of the two highly unfavorable factors to be even more unfavorable (or closer to zero), the result (0.28) is greater than either 0.1 or 0.2. It is interesting to note that had we not abandoned the multiplicative applicational model, we would have obtained a more desirable result: $p_t = 0.1 \text{ x } 0.2 = 0.02$.

This leads us to the conclusion that each model functions satisfactorily in a certain area. The multiplicative applicational model is appropriate for the joint effect of unfavorable factors, whereas the multiplicative non-applicational model is appropriate for favorable factors. In any case, we have not yet found the function we are seeking because, when confronted with real linguistic data there is no rigorously objective method which allows us to decide which model to use. Furthermore, even if there were some way to decide in each specific case, the prospect of continually switching models in an *ad hoc* search for `good' results is unacceptable.

In 1978, Pascale Rousseau and David Sankoff attempted to resolve the contradictions of the three previously described models by proposing yet another mathematical model, one that would retain the desirable properties of the other three, and yet is capable of replacing them in the analysis of any set of data. This model, known as the *logistic model*, had already been used by statisticians in other areas, such as biology, where relationships between the relevant categories are similar to those found in empirical linguistic research. In principle, the logistic model has no theoretical basis and in fact, is described as an `analytical convention' by the authors.[5] However, as shall be seen, it does have some justification because it has the advantage of enabling us to isolate the individual effect of each factor.

All of the mathematical models, at least in their simple form, deal with the occurrence of two alternatives which are in competition in the environments described by the various combinations of the p_i. If one of the alternatives has some sort of advantage in the evolutionary struggle for survival (for example, if its psycholinguistic processing is simpler), this fact should result in its being favored over the other form during the on-going process of change. In other words, at any point in time the future use of the favored alternative is more probable than that of the other because of its evolutionary advantage. With the passing of time the form which is not favored will completely disappear from the language.

According to the theory of evolution, when there are two alternatives in competition, and one of them has an advantage that makes its future use more probable, the loss of the disfavored form takes place over a period of time in accordance with a curve in the shape of an elongated *S*. In the beginning stages of the process, the favored form occurs only infrequently, and its introduction is slow. With the passing of time, the process speeds up, only to slow down again when it is near its final stage and the favored variant occurs nearly 100% of the time. This means that the increase in frequency of the new form (e.g. from 5% to 10%) is much slower than the increase from 50% to 55%, which in turn, is much quicker than the increase from 90% to 95%, when the change has slowed down again. This theoretical picture has been confirmed in empirical studies of several linguistic phenomena, such as the increase in use of the definite article before possessives in Portuguese (Oliveira and Silva 1982) and the increase in the use of periphrastic *do* in English (Kroch 1983). The empirical justification for the logistic model lies in the fact that it corresponds to an *S* curve when graphed. It seems that the logistic model performs well in the description of synchronic variation because variation is an instantaneous cross section of a process of change.[6]

The logistic model is:

$$\frac{p_t}{(1-p_t)} = \frac{p_0}{(1-p_0)} \text{ x } \frac{p_1}{(1-p_1)} \text{ x...x } \frac{p_n}{(1-p_n)}$$

In the above formula, p_t is the overall probability of applying a certain variable rule in the environment determined by the occurrence of one factor from each group. P_0 is an INPUT probability, which corresponds to the grand mean of the application of the rule, adjusted in accordance with the model. In principle, the INPUT measures the general tendency of use of the variant under examination, abstracting away from the effect of the factors.

Since the logistic model, unlike the previous multiplicative models, does not follow the probabilistic definition of independent events, the use of the term *probability* is no longer strictly justified for p_i. For the logistic model, *relative weight* seems more appropriate, though many linguists still prefer the original terminology.

The empirical data, together with the mathematical formulation of the logistic (or additive) model are not sufficient to determine unique values for the p_i. In order to do this, it is necessary to impose certain conventions, which, though arbitrary from the mathematical point of view, do make it possible to calculate p_i. The basic problem is similar to that of finding values for *x* and *y* that satisfy the equation $x + y = 5$. There are an infinite number of correct solutions, such as $x=2$, $y=3$ or $x=1$, $y=4$, or… In order to fix unique values we adopt the convention that the mean of the values for all the factors of each factor group must be equal to 0.5. In older versions of the computer program, this mean was not weighted, (i.e. each factor in a group had the same weight). In more recent versions, the mean is weighted by the number of empirical data available for each factor so that factors with small numbers of tokens will not have undue influence in the calculations.

When the logistic relative weights associated with the factors are close to ZERO, this model is very similar to the multiplicative applicational model very closely since the denominators $(1\text{-}p_i)$ of the terms become nearly equal to 1. Under these conditions the typical term $p_i/(1\text{-}p_i)$ becomes $p_i/1$, which is equivalent simply to p_i. When the weights are close to ONE, the logistic equation approximates the multiplicative non-applicational model since in this case it is the numerator that becomes negligible. Halfway between these two extremes, with weights close to 0.5, the equation behaves like the additive model. In this way, the logistic model is a synthesis of the three earlier ones. The interpretation of the relative weights using this model can be understood in the following way:

By replacing the right side of the formula by p_R, we obtain:

$$\frac{p_t}{(1\text{-}p_t)} = p_R$$

Solving for p_t we obtain:

$$p_t = \frac{p_R}{(1+p_R)}$$

Here p_t measures the chances for a variant to be realized in a given environment. As p_R increases, the fraction on the right hand side of the equation tends toward the value ONE (i.e. the rule tends toward categorical application). As p_R decreases, the fraction tends to zero (i.e. the rule tends toward categorical non-application). p_R as a product of terms takes the following form:

$$p_n / (1-p_n).$$

Thus, p_R increases for each such term with a value greater than ONE, and decreases for terms less than ONE. These terms are greater than ONE when p_n is greater than 0.5, and less than ONE when p_n is less than 0.5. When p_n is exactly 0.5 the term becomes ONE, and has no effect at all on the product. For this reason, the weights calculated in accordance with the logistic model are usually said to be favorable to the application of the rule when they are above 0.5, unfavorable when below 0.5, or neutral, when equal to 0.5. However, it should also be underlined that the numerical solution to the logistic equation is to a certain extent, arbitrary, as has already been seen in connection with the weighting of the numerical values of the factors of each group and the choice of p_0 as a corrected grand mean. In principle, the absolute values of the relative weights have no analytical significance. What is important is their order,. and precisely for this reason, the term *relative weight* is preferred. The relative weight associated with a given factor can even be 0.4 under certain mathematical conventions, and 0.6 under others, but the relative order of the values within a group will remain the same. Consequently, one must be careful before basing any conclusions on the fact that a relative weight may turn out to be less than 0.5, and thus appear not to favor the application of a certain rule. Even more care should be taken in the analysis of numerical results calculated for different data sets.

When the data for noun phrase agreement discussed above was fitted to the logistic model, 0.66 was the relative weight obtained for the simple opposition, and 0.84 for the complex opposition in the morphological opposition group. These results confirm the initial hypothesis, and are in agreement with the frequencies for the non-biased data of positions 2 and 3. They were calculated using the complete set of data, including the data from position 1. This constitutes a practical demonstration that the logistic model is capable of overcoming difficulties of analysis caused by the uncritical use of raw frequencies.

The methodology of variation theory constitutes a powerful and reliable tool that can be used for the study of any variable phenomenon at any level of linguistic activity. Its limitations are those of the individual linguist, who evidently must shoulder the responsibility of discovering the relevant factors, of correctly obtaining and codifying the empirical data, and, above all, of interpreting the numerical results within a theoretical framework. The progress of linguistic science depends not upon the numbers themselves, but upon their analysis, and how the interpretation of analytical results can contribute towards our understanding of the phenomenon of human language.

Notes

[1] This is based on the introductory part of Naro & Votre 1980. I am grateful to Maria Marta Pereira Scherre for her comments on this version. Whatever errors remaining are my responsibility.

[2] In the pairs dependent/independent and interaction/independence the sense of `independent' is not the same.

[3] An alternative would be to use simple categories and an appropriate statistical measure of the margin of error entailed by this decision.

[4] In two successive throws of a coin there are four possible results: heads-heads, tails-heads, heads-tails, or tails-tails. If the coin is `honest', i.e., non-biased, each of these result has equal chances of occurring. For any given result, then, its probability of occurrence is 1 in 4, or 0.25 and its probability of non-occurrence is 3 in 4, or 0.75. The sum of the probabilities of occurrence and non-occurrence is 4 in 4, or 1.

[5] Compare this state of affairs with the justification of the multiplicative model by reference to the criterion of independent events.

[6] This does not mean, of course, that all synchronic variation represents a dynamic on-going change. It is not at all unusual for the cross-section to be a static one, without movement in any particular direction.

References

Braga, Maria Luiza. 1977. *A concordância de número no sintagma nominal no*

Triângulo Mineiro. Masters dissertation, PUC/RJ.

Cedergren, Henrietta and David Sankoff. 1974. Variable rules: Performance as a statistical reflection of competence. *Language* 50.333-55.

Kroch, Anthony S. 1983. Function and grammar in the history of English periphrastic *do*. *Language change and variation*, ed. by Ralph W. Fasold and Deborah Schiffrin, Current Issues in Linguistic Theory 52.133-172. Amsterdam/Philadelphia, John Benjamins Publishing Company.

Labov, William. 1969. Contraction, deletion, and inherent variability of the English copula. *Language* 45.715-62. [Revised version, with no mention of the additive model in Language in the inner city, 65-129, Philadelphia, University of Pennsylvania Press, 1972.]

Naro, Anthony J. and Sebastião Josué Votre. 1980. *SWAVA -- Sistema SWAMINC/VARBRUL: Manual do usuário*. Rio de Janeiro.

Naro, Anthony J. 1981. The social and structural dimensions of a syntactic change. *Language* 57.63-98.

Oliveira e Silva, Giselle Machline de. 1982. *Estudo da regularidade na variação dos possessivos no português do Rio de Janeiro*. Ph.D. thesis, UFRJ.

Rousseau, Pascale and David Sankoff. 1978. Advances in variable rule methodology. Linguistic variation: Models and methods, ed. by David Sankoff, 57-69. New York: Academic Press.

Scherre, Maria Marta Pereira. 1978. *A regra de concordância de número em português*. Masters dissertation, PUC/RJ.