3 The reliability of dialect data

3.0. Introduction

The distinction between *reliability* and *validity* is discussed in the first sections of the previous chapter. The remainder of that chapter elaborated on validity with respect to the quality of phonetic dialect transcriptions and on the quality of the dialect data from our survey against the background of the bio-social features of the dialect speakers involved, and the differences, if any, in speech style that may influence the linguistic behaviour of the speakers. In the following chapter, 4, we will also pay attention to possible interaction effects between dialect speakers and the fieldworkers involved.

The validity of the greater part of the dialect transcriptions in our project was established. As regards the quality of the dialect data, the question whether the data represented the inherently variable "Grundmundart" got a positive answer as well.

However, an instrument may measure in a valid way - that is, measure what it is meant to measure - but at the same time do so unreliably. Compare the example from chapter 2: a thermometer measures temperature and not velocity or humidity, and as such its measurements are valid for temperature, but the instrument readings may float somewhat erratically around the markings. In that case the instrument readings are unreliable. *Reliability* is about precision, consistency of the measuring instrument. Note that the fact that an instrument measures, reliably and consistently, the same value but not the real value, produces invalid results and thus concerns validity. In this chapter we discuss two aspects of the reliability of our data.

The first aspect relates to the reliability of phonetic transcriptions. In most dialect research situations these are the format of the data that is dealt with. The question then is, whether the transcriptions are precise. We cannot answer that question in an absolute sense. There will always be mistakes, but one may evaluate the question in terms of degrees of 'less' and 'more'. Sound practice is to optimise the transformation process from raw dialect data (recordings) to research data (transcriptions). This evaluation is the subject of section 3.1.

The second aspect concerns the reliability of dialect data in a geographical

sense and is treated in section 3.2. Dialect features from the 'Grundmundart' represent area features and therefore show particular geographic patterns. The question is whether the data represent the geographical pattern in a sufficiently precise manner. Geographically bounded data in general show peculiar problems which arise from their area character.

3.1. The reliability of phonetic transcriptions

This section deals with the reliability of the transcriptions. In the following, we discuss *concordance* between transcribers in the sense of 'reliability', not in the sense of *agreement*.[1]

The state of the art and the discussion up until 1989 is treated in Glorie (1989).[2] As in chapter 2, we compare transcriptions of the two transcribers, V and H, who produced the greater part of the transcriptions and we compare these also with the transcription by R, V's IPA-trainer. A transcriber is considered to be a measuring instrument. Compared to 'real' phonetic measuring instrumentation, the human ear is in many respects unsurpassed in its resolution capacity for speech.[3] From the degree of concordance between transcribers we get information about relative (not absolute) reliability of the transcriptions produced. The degree of reliability is summarised by a reliability measure *r*. The formula is (Vieregge 1985, 170-171):

$$r = 1 - \frac{\Sigma_i^p a_i}{N \,.\, \bar{a}max}$$

where a_i is the difference per segment between two transcriptions. These differences are measured as the perceptual distance between any two segments. These distances are summed over i = 1...p differences. N is the number of vowels, consonants or diacritics. In this study, these segment classes are treated separately; āmax is the average maximal distance within the segment classes of vowels, consonants and diacritics, that is, considering the maximal distance of a certain segment from that class to all other segments from that class. This ensures that the reliability scores run from 0-1. The distances are read from a distance matrix for transcription symbols, categorised in a general system of distinctive features.

---

1 See for this distinction Cucchiarini (1993).

2 Later points of view are given in Vieregge and Cucchiarini (1989) and in Cucchiarini (1993). While Cucchiarini (1994) does not supply a scoring method for diacritics and because she is critical of the amendment to Vieregge's reliability measure, in Vieregge and Cucchiarini (1989), I use the original definitions from Vieregge (1986). These were also used by Glorie (1989).

3 Illustrated by the phonetician B. Schouten in a conference paper on /a/ realisations from Utrecht to Amsterdam and the transcription and perception of these. He compared trained transcribers' data and perception data with F1/F2 measurements and, in the case of certain mismatches, yielded to the human ear.

Ultimately, this system is based on articulatory similarity judgements. The distance matrix is thus validated articulatory. For example, the distance for the transcription difference [a]-[y] is 8; for the difference [i]-[I] it is 1. All transcriptions were made in the IPA-version codified in Chapman (1983). In this system an anomaly of standard IPA is solved: standard IPA does not provide a lowest central vowel. The two recent IPA revisions after the Kiel convention did not remedy this situation. As a consequence, one has to transcribe the lowest central vowel, if any, by front lowest plus back diacritic, or by lowest back plus front diacritic. In Chapman's system, [æ] is the lowest (open), frontal unrounded vowel, [a] is lowest (open), central and unrounded. Also [^] is half open, central and unrounded. [I] and [Y] are frontal, between closed and half closed; they are unrounded and round respectively. Symbols not occurring in Vieregge table were inserted into the arrticulatory table where they belonged because of their (distinctive) articulatory featural make-up. In the case of diphthongs, the two symbol segments were counted as two different vowels, indicating beginning and end of the glide. Diacritics in IPA, not supplied for in Vieregge's table, were fixed on the value 0.5 (this is the average value for diacritics that did have a proviso in Vieregge's system. A missing or an extra consonant was valued as 4.11 (The average of all possible deviations between consonants without diacritics). A missing or an extra vowel was set on 4.61 (the average of deviations of all vowels without diacritics). Missing or extra second diphthong elements were scored like the vowels. Scoring of [I] and of [e] was provided for in agreement with their IPA position (Vieregge gives them the same height and frontness and to make a distinction one would need diacritics. Like Vieregge did, the dimension front back was doubly weighted, but the diacritics on this dimension were not. Vieregge did not give an āmax for diacritics. āmax =3.74 was calculated from the transcription data. In this way, the average maximal distance with diacritics is less than the ones for vowels and consonants (resp. 4.61 and 4.11). This value 3.74 is close to the maximum possible difference for diacritics (4.0). All Glorie's r-values were recalculated by me, because Glorie's data contained possibly input-mistakes. These were the consequence of using a calculator instead of a spreadsheet (see also Glorie 1989: 41).

The matrices underneath are given with r-values between all pairs of transcribers for vowels consonants and diacritics apart. I am confining myself to the comparison of V and H with R because they produced most of the transcriptions en because the two had a different IPA-background. V got his phonetics training from R, H did not (cf. chapter 2). A reliability analysis of all transcribers is incredibly time consuming. First of all, every transcriber has to transcribe the same fragment (say Egmond aan Zee), then one has to score the transcriptions. The number of pairs of transcibers to compare grows fast.[4] We will see that V and

[4] Ultimately, and for the number of transcibers n, one has a triangular matrix of r-values counting: (n-1)*n/2. For 6 transcribers, this amounts to 15 pairs. If every word from the list should

H were very good, but the reliability of the other transcribers is guaranteed by the fact that most of them had their training during their professional education as linguists from R and only the ones who got the mark "very good" in their final tests could participate in the project. The requirements for participants from other 'schools' were high. New transcribers were coached at the beginning and supervised at a later stage by the experienced transcribers in the project. If necessary, the maintenance of the 'standard' was effectuated in personal contact with R. As was the case with V and H, who showed remarkable convergence by regular contact with each other, contact of other transcribers with V, H and R did lead to the sort of convergence that is a prerequisite for reliability. For reliability analysis we used the tape of the locality of Egmond aan Zee from the GTProject originally transcribed by H. Later, R en V agreed to transcribe 200 items on the list. These 200 items were analysed. R-values are given following two methods, the first according to Vieregge (1985), while the second follows the procedure of Almeida and Braun (1986). Almeida and Braun use a constant a-max of 3.0 and take N to be the maximum number of segments per transcriber (N_{max}). Vieregge fixes N to be the average number of transcribed segments (N_{mean}). Almeida and Braun's scoring of transcribers' deviations was not used, because their scoring is not calibrated against perception data[5] as has been done for Vieregge's. We used Vieregge's scorings instead.

R-values according to Vieregge

Vowels without diacritics

V	—		
R	.909	—	
H	.929	.900	—
	V	R	H

Concordance is highest between V and H, the two fieldworkers who produced the largest number of transcriptions in the project. Second highest is the concordance between R and H. R is V's IPA-phonetics teacher, but not H's. Therefore, one would expect R and V to maintain more ore less the same transcription standards. However, during the fieldwork and during transcription H and V discussed problematic cases with each other frequently. One may expect them to show a very high concordance too, although they are from different IPA-'schools'. Both expectations are not disconfirmed.

---

have 6 segments to compare (a rather low estimate), then every pair of transcibers would amount to a matrix of 200 words*6 segments=1200 segments. This forms a matrix of (1200-1)*1200/2 and results in 719400 individual comparisons pro pair. Analysing all transcribers from GTP-Nederland would constitute a research project by itself.

5 See Vieregge et al. (1984).

R-values according to Vieregge

Consonants without diacritics

V	—		
R	.949	—	
H	.976	.952	—

With consonants, r-values are higher than with vowels. The transcription of consonants is less difficult. The relationship in this consonant matrix is somewhat different from the situation found for the vowels: V shows high concordance with H (in this respect, there is no difference) but R accords better with H than with V. This means that R and H have sometimes opted for a different choice of basic symbol. Inspection of the transcriptions shows that these differences pertain to variants in the subclass of resonants /r/ and /l/.[6]

R-values according to Vieregge

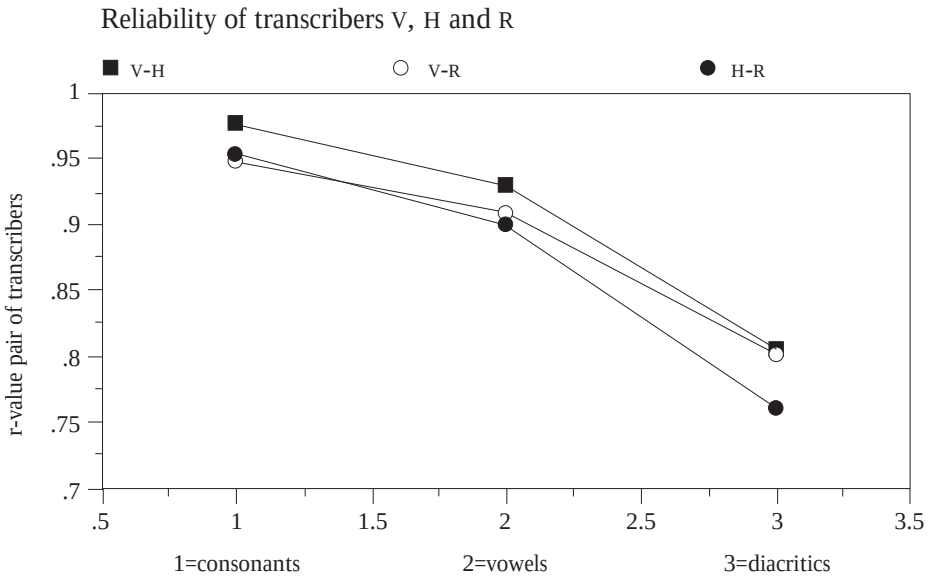
Diacritics

V	—		
R	.801	—	
H	.806	.761	—
	V	R	H

The diacritics for vowels and consonants are taken together. Concordance is again highest between V and H, and then between V and R. Concordance is least between R and H. V and R maintain a common standard even in the use of diacritics. The concordance between V and H is higher than that between H and R.

[6] In the case of /l/, the influence of R's research on the vocalisation of /l/ in regional variants of Dutch in the years before his transcription may have influenced his transcription practice (Van Reenen 1986).

Figure 1: Method Vieregge



The concordances are given in Fig. 1. The degree of concordance is highest between V and H, compared to that of each with R. Therefore, the reliability of the transcriptions produced by V and H is very high, even more so if we consider the fact that the number of segments (consisting mainly of 200 words plus a few word groups) is considerable.

R-values according to Almeida-Braun, as a result of their scoring method, show lower concordance. In this case it is best to look for indications to the extent to which trends (rankings between transcriber pairs) found with the first method remain stable in the second.

R-values according to Almeida-Braun

Vowels without diacritics

V	—		
R	.761	—	
H	.812	.731	—
	V	R	H

Consonants without diacritics

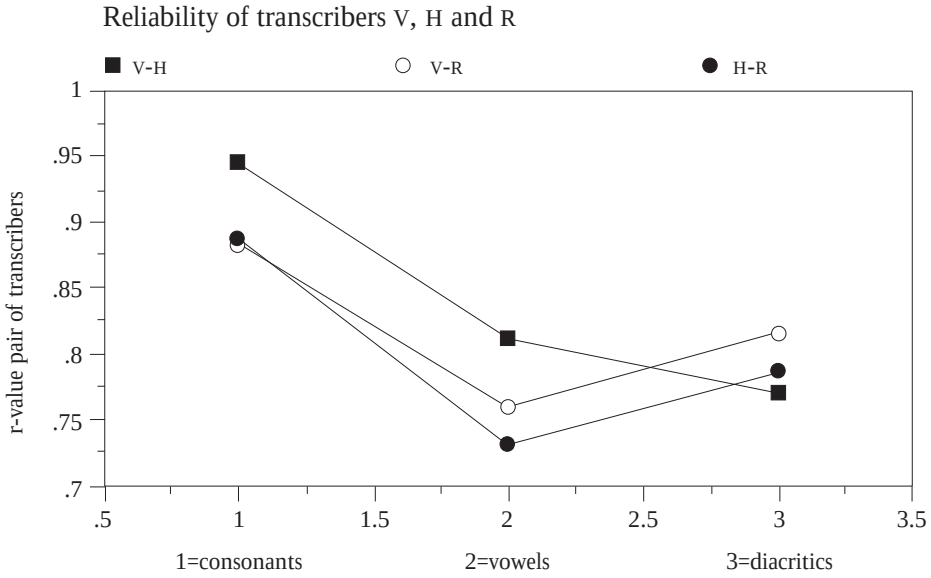
V	—		
R	.882	—	
H	.945	.888	—
	V	R	H

Diacritics

V	—		
R	.815	—	
H	.771	.786	—
	V	R	H

With vowels as with consonants, the highest concordance is between V and H. The rank between the pairs is the same as with Vieregge's method. With diacritics, on the other hand, the concordance is higher in the pair R and V. The rankings between V-H and R-H have changed, V and H both tend to a more extensive use of diacritics than R does. This tendency shows more clearly in the measurements according to Vieregge. Compare Fig. 2 with Fig. 1.

Figure 2: Method Almeida-Braun



If measured by Almeida and Braun's method, the concordances are high (see Fig. 2), albeit less high than according to Vieregge, and particularly so with the vowels. The different results of the two methods seem to be a consequence of the formula used, because the differences between pairs of transcribers remain the same (diacritics excepted) with vowels and with consonants as well. Since there are no fundamental reasons to choose the minimum of the two (N=Nmin), one may prefer the average, N=Nmean, as is the case in Vieregge's formula. The advantage of Vieregge's method is that it gives a better idea of the situation than Almeida and Braun's. On the one hand, we cannot give a definitive pronouncement on the question whether Vieregge's method overestimates reliability. On the other hand, the use of N=Nmax by Almeida-Braun definitely underestimates reliability in the comparison of pairs of transcribers.

The general conclusion is that the transcriptions produced by V and H are indeed reliable and, accepting the results of Vieregge's method, very reliable.

3.2. Reliability of traditional dialect geography. A statistical appraisal[7]

3.2.0. *Introduction*

In this section we show that traditional dialectology can produce reliable and consistent data; the reliability of dialect data with respect to their regional aspects.

I assume that the traditional opinion on dialect as "Grundmundart" is valid. Therefore, the population to sample from are the speakers of these varieties. The sample procedure often used by traditional dialectologists is a form of purposive quota sampling (Werlen 1984, 100). Sampling in *quota* means that the sample is obtained by the application of biosocial criteria such as age, sex, educational achievement, socio-economic status, mobility, in order to obtain homogeneity (ibid. 102-104).[8] *Purposive* means that the knowledge about the distribution of these criteria is used in the sampling procedure.

We have seen in chapter 2, sections 2.3 and 2.4, that such a homogenisation is not necessary, let alone desirable. This sampling procedure has a disadvantage compared to simple random sampling; one cannot infer beyond the sample to the population of dialect speakers because the sampling error cannot be computed. This is the classical view. The classical view on the interpretation of inferential tests (tests of significance) presupposes that one is sampling from a completely specified list that enumerates the whole population under consideration in order to infer from the sample to the population. This means in the case of dialects that we would possess a list of all the dialect speakers to sample from. I believe that such a list belongs to the world of the nearly impossible.[9] The conclusion is that this procedure cannot be used. As to the geographical aspect of dialects, the procedure would imply that we had an exhaustive listing of all the localities where dialect in the sense of "Grundmundart" is spoken to sample from. Such a procedure is as impossible as the former and cannot be used.

Application of tests of significance is possible if we are willing to make some additional assumptions with respect to the dialect sample we have. There are several possibilities, which differ in accordance with the aspects of the problem that are accentuated. The central problem is that we want the error component in the data to be as small as possible compared to the component that could explain the variation in our data. Henkel (1976, 85-87) mentions three aspects:

A. The sample is a random sample from some assumed, conceptual universe. The research question that can then be answered is: could sampling error account for any of the differences between expected and observed values? The expected values are considered as standing for the universe.

---

[7] This part is a revised version of Goeman (1986).

[8] Altmann and Naumann (1982, 654-656) give other sampling procedures, including stratified ones.

[9] However, in some rare cases an estimate of this population of dialect speakers can be made, as was seen in chapter 2 section 2.3.2.

B. Random measurement error is the source of randomness in the data. Research question: could random measurement error account for the differences found? If unexplainable as mere error, the results are due to an important causal factor. In our case this factor is spatial distribution.
C. The data are generated by some random process. Research question: could the results have been produced by chance? If not, some causal, theoretically important factor accounts for the negative result that random factors did not produce the observed value: in our case, the geographically distributed values.

The two last aspects, B and C, are at the basis for what follows. Aspect B, concerning measurement error, is developed in section 3.2.4 and is directly connected to the question of reliability. We sample the same population (the population of dialects) twice to see whether the results are comparable and to which degree. If the outcome is positive, reliability is assured. We develop aspect C, the random process, in section 3.2.3, to show the importance of the geographical factor.

Dialectologists have always considered their data to be representative for the dialect spoken in the localities under consideration; they consider their data as reliably representing those dialects. This position is assumed here in the following framework: a) dialects are not homogeneous, they show variation, and b) processes show regional differences. We want to get an idea of the reliability of traditional dialect data in the geographical sense: to what extent dialect data are consistently explained by regional characteristics.

As an illustration we examine a specific variable which has received considerable attention over the last 25 years, the variable t-deletion,[10] in section 3.2.1. In section 3.2.2 we explain the specific assumptions of the model. Then we elaborate on aspects B and C in sections 3.2.3 and 3.2.4.

3.2.1. *Regional trend in t-deletion*

T-deletion can be scored as absence (1) or presence (0) of word-final t, but also as present, but unreleased (0.5). Regional effect was measured by the geographical co-ordinates for the locality where the dialect was spoken. There were 52 localities in the sample and 116 potential instances of t-deletion per location, thus N=6032.

Regional trend analysis[11] was adopted by Goeman and Van Reenen (1985)[12] as a descriptive instrument to separate a general trend in t-deletion from residuals indicating strictly local phenomena concerning t-deletion. Regional trend analysis belongs to the class of multiple regression models. In our case, the percentage

---

[10] Data from the GTP-project. The data for this section were collected in project period 1979-1984. The phonetic transcriptions were then available on a PDP 11-70. The fieldworkers in this period were J. Aben, J. Buitenhuis, D. Coppes, L. Gijsbers, A. Goeman, A. Ottow, P. van Vliet and C. van Zaanen. They also made the transcriptions of the taped sessions.

[11] For this technique see Haggett (1970), and Hodder and Orton (1976).

[12] In this study chapter 6, section 6.4.2.2.

of t-deletion was seen as partly explainable by the two geographic dimensions north-south (NS) and west-east (WE). We adduced tentative substantive extra-linguistic explanations for the region that showed high positive residuals. Connections were established with the relatively isolated position of this region (the Betuwe).

Goeman and Van Reenen employed a linear function because it was simpler, more restricted, therefore more vulnerable to violation by the data, and also because square and cubic functions (being non-linear) are not easily interpreted in terms of distances on a dimension (see for results table 3.1). The exposition in this chapter being rather technical, the reader may consult chapter 6, section 6.4.2.2 at first, and then chapter 4, to get acqainted with the basics of regression models applied as trend surfaces.

The coefficient of determination R^2=0.056, so a little less than 6% of the variation is explained by the geographical co-ordinates in all. The only relevant dimension seems to be the direction west-east (WE): r^2=0.227, so 23% of the variance is explained by geographical position along this dimension, and 77% of the variance remains unexplained. This is not to say that this 77% is all error, on the contrary, a large part of it is possibly explainable by other factors, such as socio-economic factors and dialect-internal factors of a structural linguistic nature. The former, extra-linguistic factors are the subject of chapters 4 and 5; chapter 6 is concerned with the latter, structural and internal linguistic factors.

It was also noted by Goeman and Van Reenen[13] that variances seemed to increase going from west to east (as can be seen from the residuals in map 1). It was suspected that these larger residuals represented variance that was not constant (hetero-scedasticity). It seemed not useful, therefore, to test for statistical significance. In this part I want to investigate the violation, if any, of this assumption as well as the violation of other assumptions about the linear model by the data.

3.2.2. *Assumptions of the model*

The multiple regression model used in this chapter has the form:[14]

$$Y_i = \alpha_1 + ß_1 X_{1i} + ß_2 X_{2i} + \varepsilon_i \quad \text{for the population, and}$$
$$TDEL_i = a_1 + b_1 X_{1i} + b_2 X_{2i} + e_i \quad \text{for the sample,}$$

where a_1 is an estimate of the mean value of TDEL if b_1X_1 and b_2X_2 are to be 0.

TDEL is the dependent variable t-deletion, the value of which is a function of the independent variables; X_1 and X_2 are independent variables and stand for the two geographical co-ordinates NS and WE (i.e. north-south and west-east), *e* is the error component that is the difference between the estimated value of TDEL and

---

[13] In this study chapter 6, footnote 9.

[14] Chapters 4 and 5 contain extensions of this regional trend model with other variables.

the value as found in the data. Thus the error is the difference between model and data. Therefore $\{a_1 + b_1 X_{1i} + b_2 X_{2i}\}$ is the systematic, explained variation in the data, $\{e_i\}$ is the random, unexplained variation in the data. Naturally, we prefer this error to be sufficiently small. If it is too large, the data may be unreliable in a geographical sense. *i* is a subscript that ranges over the localities.

The assumptions of the multiple regression model are:[15]

1. The model is linear.
2. a. The X's are fixed (non-stochastic: $E(X_{ij} e_j)=0$).
   b. The X's show no exact linear relationship: (no multi-collinearity).
3. a. The error has 0 expected value: $E(e_i)=0$.
   b. The error has constant variance for all observations: $E(e_i^2)=s^2$ is constant; constant variance is homoscedasticity.
   c. Errors of observations are not correlated: $E(e_i e_j)=0$ for i j: no serial or autocorrelation.
   d. Errors are normally distributed.

Another assumption is about the completeness of the model. Completeness signifies that all relevant independent variables are specified in the model. This assumption is treated in chapter 4.

Assumptions 1, 2a, b and 3a are justified in our case: the model is indeed linear, the geographical co-ordinates are fixed, the geographical co-ordinates are independent of each other and there are no reasons to think that the mean value of the error would not be about 0.

Assumption 3d is important for tests of significance, but when the sample size is large enough, violation of it is not serious, because normality is approached when sample size increases (Lewis-Beck 1980, 30).

Parameter estimates remain unbiased and reliable if 3b and/or 3c are violated, but in those cases significance tests and confidence intervals are less appropriate because they may underestimate the real values.

As mentioned above, Goeman and Van Reenen (1985)[16] assumed that the data showed heteroscedasticity due to increasing variance. T-deletion was expressed in the form of percentages and we expect that percentages, like proportions, have unequal variances. The increase in variances could partly be an effect of the properties of proportions; not only are they count data and not integer scores, but they are also bounded by 0 and 1 in the case of proportions.[17] The first step in solving this problem is transforming the data.

Transformation of the data is all the more necessary because the counts for unreleased /t/ are very infrequent an d we are left with a variable taking on rather

---

[15] See a standard handbook like Pindyck and Rubinfeld ²1982, 76.
[16] In this study, chapter 6, note 9.
[17] Percentages are simply reexpressions of proportions in the 0-100 domain.

extreme values in the [0-1] interval. This means that large changes in the independent variable are needed to provoke small changes at the extremities of this [0-1] interval, while normally the influence of the geographic variables on t-deletion is largest when TDEL has values of around proportion=0.5. Therefore the logit probability model was chosen as a transformation of the proportions of the dependent variable TDEL: $\mathrm{Ln}(\mathrm{Prop}_i/(1-\mathrm{Prop}_i))$, see Pindyck and Rubinfeld ibid., 287-291. This makes variances more constant.

After transformation of the data we tested for heteroscedasticity in the data: the remaining errors were correlated with north-south and west-east separately by means of Spearman's Rank Correlation Coefficient.[18] The correlations were not significant in either case (table 3.2), and the assumption of homoscedasticity can be upheld.

Serial or autocorrelation could be a problem too. Because our data are geographical data we could have spatial serial or autocorrelation. That implies that the scores of for example four locations A, B, C, and D surrounding the location Z, affect the score of that location Z: locations with low rates of t-deletion will have locations with low(er) rates neighbouring them, and localities with high rates will tend to have high(er) rates in their neighbourhood. The dialect-geographical analogue of this is area formation by dialect features. Geographical data normally show spatial autocorrelation characteristics.

The Durbin-Watson test for serial correlation is appropriate only when cases can be arranged in one direction. It is not applicable to geographically arranged data, because localities from different directions exert their influence on location Z.[19] The west-east dimension seems to be by far the most important one in explaining the variance, and therefore it is possible to get an indication whether autocorrelation is present, by drawing an imaginary line west-east and by taking all the localities on close to this line, and then computing Durbin-Watson's d. This test now results in the value: $d=2.49162$ ($2<d<2.59$), which means that serial correlation does not exist in this case.

It is possible, though, to interpret the geographical distance among the dialects along the west-east dimension in terms of the distance between these dialects and those dialects in the west of the country that resemble the standard language (see Daan and Blok 1970, 9 and 32-33 and map colours indicating this distance). In terms of these uniformly oriented distances, applying the Durbin-Watson test does make sense. For that reason result of the test is given between parentheses in all relevant tables in the appendix.

We may conclude that the assumptions for the linear model employed are met. Another result is that, given position C concerning tests of significance, we

---

[18] In the regression, Ordinary Least Squares (OLS) is used as an estimation method and thus the correlation between the error and each of the independent variables is always zero (Koutsoyiannis [2]1977, 185). Therefore, Kendall's correlation coefficient is not appropriate in this case.

[19] At the time of writing this article, the computer program for testing for spatial autocorrelation (Hodder and Orton 1976, 179) was not available to me. Standard statistical packages like SPSS do not provide these tests. Spatial autocorrelation is treated in chapter 5.

are justified in using these tests, and we are justified in computing confidence intervals. As the significance test turns out to be significant, the data are not generated by a random process, and therefore the theoretically important factor of the dialect's geographical position, as given by its geographical co-ordinates, produced the observed t-deletion values. In the following section we will see that this is the case.

3.2.3. *Are our fieldwork data generated by some geographically random process?*

The result of the F-test for the whole function that describes the relationship between geographic position and t-deletion, is significant ($F(2,49)=8.16796$; sign. $p = 0.0009$). Therefore, the question "are the data generated by some random process?" can be answered in the negative. Two of the function's three parameters: the intercept and the west-east dimension are significantly different from 0 as measured by the t-test, as can also be seen from the 95% confidence limits (see table 3.3). 25% per cent of the variation in the data is explained by the two geographical dimensions together ($R^2=0.25003$). See map 2 for the direction of the general trend surface and map 3 for the residuals respectively.[20]

West-east alone accounts for 25% of the variation, as can be seen from the semi-partials given in table 3.3. These semi-partials show the effect of one independent variable when the effect of the other independent variable is neutralised. When squared and multiplied by 100, it gives the unique contribution of this independent variable to t-deletion in percents (labels on arrows[21]).

It is to be recalled here that the position taken about significance testing in the case of non-random sampling is methodologically perfectly legitimate. Opinion, however, differs on the question whether denial of the constellation (in the sample) where the data could be generated by a random process, corresponds meaningfully with the constellations in the real world (the dialect population).

Yet, even in the case of simple random sampling from a well-specified population, the generalisations made are only valid for the population sampled from. The "real world" that is delimited by one's theory can be larger than the population used to test this theory. That is, theoretical claims about language variation and language change tested by means of a simple random sample in an "urban dialectological" or sociolinguistic study are not generalizable beyond this limited

---

[20] The general trend surface is a plane (running as best as is possible through the main direction of the cloud formed by the transformed TDEL-scores) the tilt of which is given by the parameters of the function for the (two) geographical co-ordinates: -.16397 for NS and +0.28877 for WE (table 3.3). The plane contains the predicted values for TDEL. See chapter 6.4.2.2 for an explication of trend surface and residual. In that section, the interpretation of the residuals as indicators for a specific local constellation, separated from the general trend, and not as merely disturbing error is stressed. The point is connected to the aspect of autocorrelation in the data. This point will be elaborated in chapter 5.

[21] (-0.14663)2 *100 = 0.02150*100=2.15; (0.49946)2*100 = 0.249460*100=24.95.

population, unless one is willing to take the same position concerning significance testing (A or B or C above) as is taken here.

To be more certain about the meaningfulness of the correspondence of the results found above with the real world, we construct a sample from another world, like the real one, so that we can know for certain that t-deletion is distributed by a random process: a permutation is made of the real locations over the geographical positions. Locations have been reallocated to each of the geographic dimensions separately by consultation of a random number table.

The results are in table 3.4: the F-test for the function is not significant, explained variation is minimal (only 8% and much lower than in table 3.3), the estimated values of the geographic parameters for north-south and west-east are not significant by t-test either. This means that the importance of the supposed theoretical factor (the geographical one) is real and that the results have not been produced by any random factors. We conclude that our data are geographically reliable.

3.2.4. *Do geographical factors cause reasonably stable patterns over time? A comparison of two samples independently drawn from the same population*

Measuring the temperature twice gives an idea of the precision of a particular thermometer as a measuring instrument. This means that replication of the measurement and the comparison of original and replicated data is a possibility to assess the reliability of the measuring instrument. If the two measurements correlate highly, the instrument is reliable, if they do not, the instrument is unreliable. In section 3.1 this procedure was used in the assessment of the reliability of transcriptions between transcribers.

We can apply the same procedure in dialectology. Because of lack of funds this procedure is seldom used. However, it is possible to make a comparison with data from earlier surveys that share the same assumptions. There is a possible catch here, because the comparison could be biased by the time-depth of the earlier data. For example, the dialect of Zoetermeer had different proportions of standard-language realisations and different dialect realisations for lower frequency and higher frequency words. However, both word classes showed a different but parallel development over time in the same set of words (N=48) as measured in 1895, 1932, 1958 and 1979 (Goeman 1984, 121-126). This means that dialect behaviour shows a certain stability between word classes over time, in this case the real time of nearly a century. To determine whether the same relative stability can be seen for regional dialect data, stability here interpreted as an indication of reliability, we compare our fieldwork data to those from the Reeks Nederlandse Dialectatlassen. The latter data differ from the former in that they containi clauses translated into dialect, instead of single lexical items. Therefore we expect that the rates of t-deletion will be higher in the RND-data.

Another difference is that the RND has a net of localities that is much denser than ours, but has fewer lexical items.[22] Thus, proportions in the GTP data will be more stable.

I singled out the Betuwe area, because the RND-data from this region were analysed by Van Hout (1980) by means of two descriptive statistical techniques: cluster analysis and scalogram analysis. According to his study the lexical items in the RND formed *one* scale (i.e. one dimension) with respect to t-deletion and therefore we felt justified to take an overall proportion of t-deletion as expression of the dependent variable in the regional trend analyses.[23] We expect that the two methods of cluster analysis on RND data and regional trend analysis on RND data as well as on GTP data will converge in their results.

Limitation to this area has one other advantage: it permits the establishment of a regional trend in another direction. For the RND-data, intercept and north-south direction are significant. This is the trend in another direction. The F-test for the whole function is significant too; the slope of north-south is outside the 95% confidence interval. The function explains 38% of the variation (see table 3.5) which is a higher value than the one found for the GTP sample in its totality (25%).

For the Betuwe subsample of our fieldwork data (GTP, see table 3.6) the F-test is not significant, only the t-test for the direction north-south is, and this slope is also outside the 95% confidence interval limits. In this respect GTP is like RND. The function explains 28% of the variation (the whole sample was 25%). The non-significant result of the F-test is not strange: a sample size of 16 is very small (cf. SPSS-Update 1981, 114). Yet we may be confident about the results, because in both Betuwe samples only north-south is significant by t-test, the semi partials even show the same direction (negative signs) and both samples show comparable values for them: -0.61951 (NS-RND), -0.53208 (NS-GTP), -0.05064 (WE-RND) and -0.01088 (WE-GTP).

The intercept in the RND data is larger than in the GTP data, 6.96 versus 6.30 and this confirms our expectation that rates of t-deletion would be higher in the RND data.

For general trend and residuals see maps 4-6. The pattern of residuals shows even more detail than the cluster analysis results of Van Hout (1980), especially the pattern below the River Waal.[24] Apart from that, the results of Van Hout's

---

[22] RND Betuwe: 78 localities and 20 word forms per locality. Our project (same area): 16 localities and 116 word forms.

[23] In this chapter we concentrate on the stability in time. The comparison between RND and our data is reconsidered as two different points in time in chapter 5, to pinpoint possible differences. The analysis there will be based on the whole sample not only for the River area but for the province of South Holland as well and takes the results of chapter 4 into account.

[24] The non-occurrence of this pattern may be an artefact of cluster analysis which, on each new clustering level and starting from then established clusters, maximises differences between candidate clusters and minimises differences within. Regression does not force grouping into discrete units. Grijns' (1991) demonstration that Jakarta Malay consists of clearly discrete

cluster analysis and the regional trend analysis converge.

In both cases the north-south dimension is preponderant in this restricted area. Not only do the two trends resemble each other, although from different time points and from two different samples, but so do the residuals.

3.3. Conclusions

Three transcriptions of the same dialect, with the same items and by three different persons, were compared segment by segment.

Vieregge's r-measure was applied in its original form, in order to determine the reliability of most of our dialect data.

We applied the statistical technique of regional trend analysis, which separates regional from strictly local factors, as a model against which our data have been tested to get information about the reliability of the geographical distribution. We have seen that geographical location is a satisfactory explanation for a sufficient part of the variability found. With some care one can even extrapolate to a population of dialect speakers.

There are many objections to the reliability of traditional dialect data. Pickford (1956) has been the first to question the sampling procedures used for the surveys of the American regional dialect atlases. Petyt (1980,110-114) subscribes to Pickford's criticisms, reproaching dialectologists that, 25 years after Pickford's article, they are still exclusively concerned with "genuine" dialect,[25] uncorrupted by contact with other forms of language behaviour. Traditional dialect data is said to show a historical bias, and research in traditional dialectology is said to have a predilection for rural speakers and elderly men. Dialectologists nevertheless pretend to do dialect *surveys* .[26]

Pickford's main objection is that dialectologists do not draw simple random samples from the population. Certainly, this can be a problem,[27] but we have shown that traditional dialect geography can produce reliable data, because the instrument used was consistent and stable.

---

dialects could be the effect of certain properties of the cluster analysis he applied. This objection is less appropriate for the Correspondence Analysis that Grijns also applied. Correspondence Analysis is the nominal analogue of Factor Analysis; with this difference that Correspondence Analysis scores variables *as well as* cases (individuals) on the same dimensions, whereas in Factor Analysis either variables or cases are scored.

25 I have already treated the variant forms of language behaviour, historical bias, the predilection for "genuine" dialect and the homogenisation to NORMs (Non Mobile Older Rural Males) under the heading of the validity of dialect data in chapter 2.

26 So for example Orton's (1962) "Survey of English Dialects".

27 Bailey, G. en M. Dyer (1992) reject Pickford's criticism. Simple random sampling is not the only sampling method possible. They stress the fact that even in sociolinguistic research most sampling methods are not simple random sampling, most methods are stratified.

Our conclusions are summarized in two points:

1) the reliability of most of the transcriptions in the GTP database is high; highest for consonants, somewhat less high for vowels, and lowest for diacritics. But even the diacritics are comparable across transcribers, they remain consistent and therefore interpretable to a sufficient degree.
2) the European tradition with its interest in the regional distribution of linguistic features and its preoccupation with "genuine" dialect is able to produce reliable results. In this tradition dialectologists have never pretended to give a survey of the language variants of the *whole* population of certain administrative units. As shown, they made claims about the population of dialect ("Grundmundart") speakers that turn out to be justified from a statistical point of view.

Whether other factors than geographical ones co-determine dialect data as sampled in the GTP database, and if so, which of these factors are the more important, is the subject of the next chapter. Further geographical aspects, which involve specific characteristics of localities either viewed in isolation or viewed in connection with their immediate environment, are the subject of chapter 5.

Appendix

In all tables: NS=North-South (co-ordinate); WE=West-East (co-ordinate)

Table 3.1: Raw percentages

function: TDEL=-5.359-0.726NS+3.845WE
R^2=0.056

function: TDEL=-13.461+3.682WE
r^2=0.227

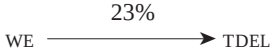


Table 3.2: Test for heteroscedasticity

Spearman correlation:	error with NS: -0.0516	t-sign.=0.359 (n.s.)
	error with WE:-0.0186	t-sign.=0.449 (n.s.)

Table 3.3: Logistic transformation of proportions

Mean substituted for missing data. * = significant

Fieldwork

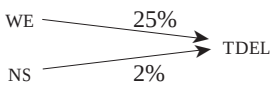
function: TDEL=-3.99891-0.16397NS+0.28877WE

t-test:	-4.301	-1.038	4.036
sign. t:	0.0001*	0.3046(n.s.)	0.0002*

R^2=0.25003, F(2,49)=8.16796, sign. F: 0.0009*

95% confidence interval contains NS=0 (slope = 0)(n.s.)
does not contain constant=0, nor WE=0(*)

semipartials: TDEL-NS=-0.14663
TDEL-WE= 0.49946



Residuals: min=-2.5625, max=4.5335, mean=0.0, stdev=1.5785, N=52
(Durbin-Watson=0.51315 (positive serial correlation)).

Table 3.4: Randomized locations

Mean substituted for missing data. * = significant

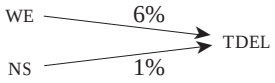
function: TDEL=-4.31683-0.05422NS+0.16336WE

t-test:	-3.893	0.240	1.803
sign. t:	0.0003*	0.8116(n.s.)	0.0776(n.s.)

R^2=0.08483, F(2,49)=2.27112, sign. F:0.114(n.s.)

95% confidence interval contains NS=0 and WE=O (slope=0)(n.s.)
does not contain constant=0.

semipartials: TDEL-NS=0.03421
TDEL-WE=0.24940



Residuals: min=-2.5097, max=4.8891, mean=0.0, stdev=1.7325, N=52
(Durbin-Watson=2.08858 (no serial correlation)).

Table 3.5: RND: Betuwe area

Mean substituted for missing data. * = significant

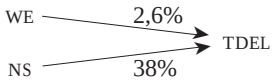
function: TDEL=6.95939-1.04577NS-0.03916WE

t-test:	4.661	-6.835	-0.439
sign. t:	0.0*	0.0*	0.6619(n.s.)

R^2=0.38640, F(2,75)=23.61474, sign. F: 0.000*

95% confidence interval contains WE=0 (slope=0)(n.s.)
does not contain NS=0 (slope is different from 0) and
constant=0 (both *)

semipartials: TDEL-NS=-0.61951
TDEL-WE=-0.05064



Residuals: min=-2.4281, max=2.6551, mean=0.0, stdev=1.0482, N=78
(Durbin-Watson: 0.53106 (positive serial correlation).

Table 3.6: Fieldwork: Betuwe area

Mean substituted for missing data. * = significant

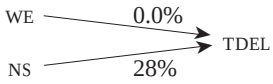
function: TDEL=6.30326-1.26707NS-0.01763WE

t-test:	1.000	-2.266	-0.039
sign. t:	0.33(n.s.)	0.0412*	0.9693 (n.s.)

R^2=0.28400, F(2,13)=2.57832, sign. F: 0.1140 (n.s.)

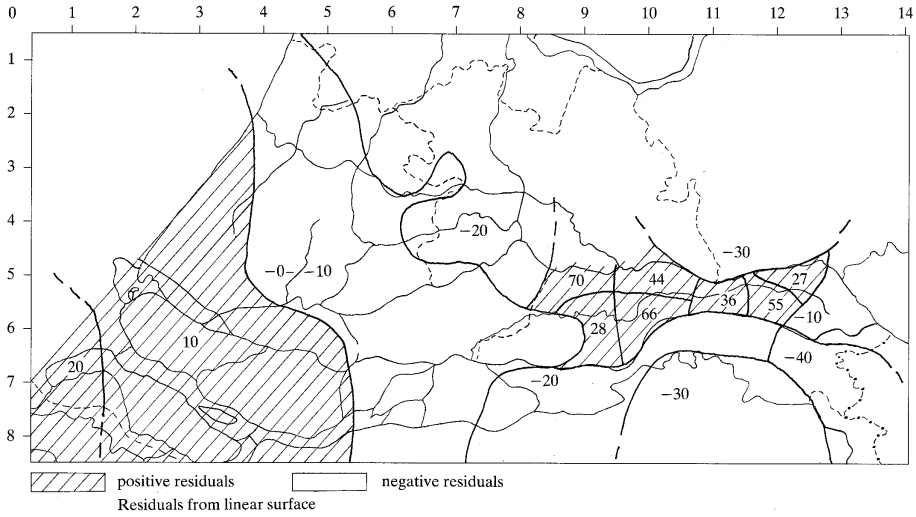
95% confidence interval contains WE=0 (slope =0)(n.s.)
contains constant=0 (n.s.)
does not contain NS=0 (slope is different from 0)*

semipartials: TDEL-NS=-0.53208
TDEL-WE=-0.01088

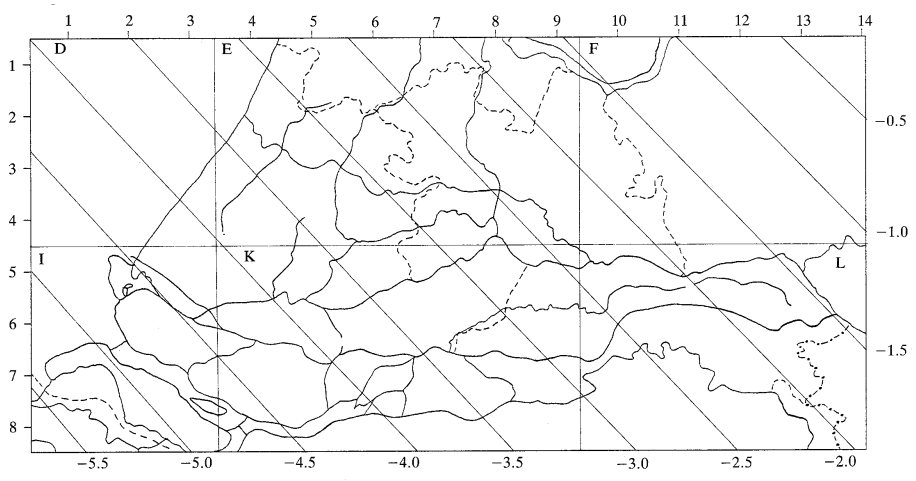


Residuals: min=-2.9942, max=3.5662, mean=-0.0, stdev=1.8873, N=16
(Durbin-Watson: 0.53106 (positive serial correlation)).

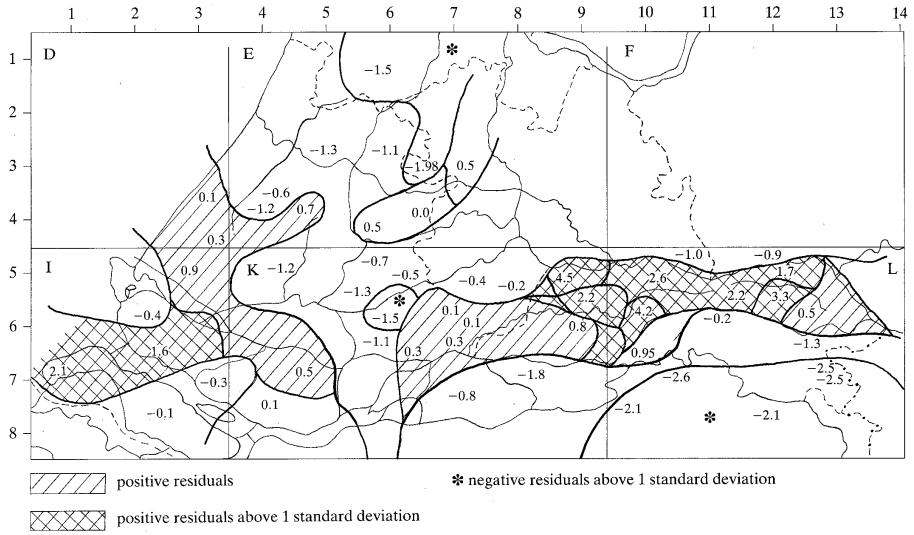
Map 1: Negative and positive residuals from trend in percentage t-deletion



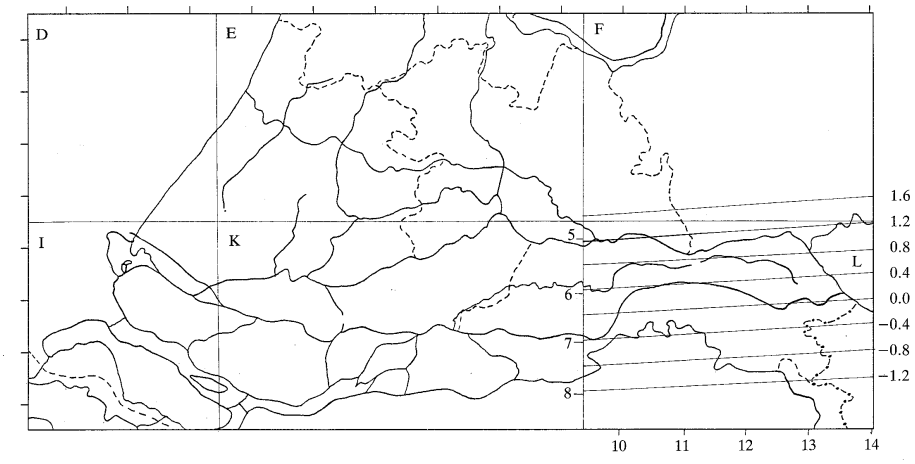
Map 2: Linear trend surface: logit of proportion t-deletion



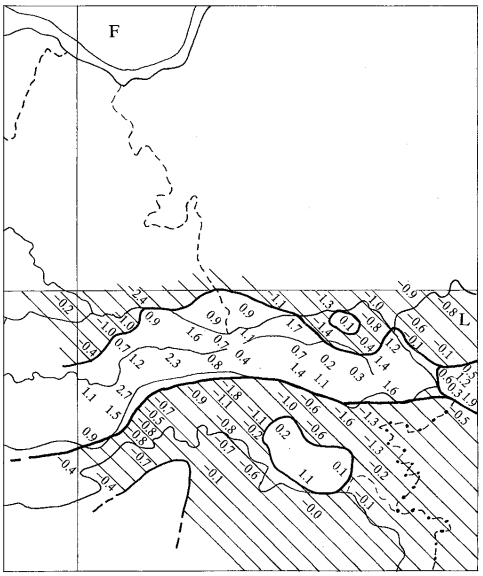
Map 3: Residuals from linear trend surface: logit of proportion t-deletion



Map 4: Linear trend surface RND-data: Betuwe-region, logit transformation



Map 5: Residuals from linear trend surface RND-data



Map 6: Residuals from RND-data exceeding 1 standard deviation