University of Pennsylvania Working Papers in Linguistics

Volume 3
Issue 1 *(N)WAVES and MEANS: A selection of papers from NWAVE 24*

Article 8

1-1-1996

Reaching Criterion in Phonetic Transcription: Validity and reliability of non-native speakers

Lisa A. Lane

Robert Knippen

Jeannette Denton

Daniel Suslak

Reaching Criterion in Phonetic Transcription: Validity and reliability of non-native speakers

Reaching Criterion in Phonetic Transcription: Validity and reliability of non-native speakers

Lisa Ann Lane, Robert Knippen, Jeannette Denton and Daniel Suslak
University of Chicago

Learning to produce phonetic transcriptions is a function of structural push and pull in which the allophonic level is relevant to understanding the transcriber's ability to reach criterion. Notwithstanding that some scholars suggest the employment of transcription by machine analysis as well as "auditory comparison", it is still not possible to associate instrumentally produced data with orthographic data and such methodology does not obviate the problem here. The data underscore that more caution must be taken in the production and especially the use of transcribed data, regardless of the language abilities of the transcriber. The most obvious question to be answered is, of course, whether or not it is possible for linguists to produce reliable *and* valid transcriptions.

1 Introduction

A critical indeterminacy of sociolinguistic transcription is the relative weight of two factors: (1) what the transcriber believes a sound should be (based on native speaker knowledge and/or adherence to a phonological theory of the language) and (2) the language independent perceptual acuity representing what is actually produced by the speaker.[1] As discussed in Nettelbladt (1993), one of the current problems with evaluating the data used in linguistic studies, is the simple fact that outside of phonetics and phonetic studies, the methodology and phonetic/phonological theories employed in determining the requirements and limitations of the specific transcribing task are rarely discussed in the publication of the study. She stresses that this practice ought to raise questions as to what the possible modifications made to the transcribing system were; what the transcribers' competence, experience and orientation to the project was; as well as what the reliability of the transcribed data is. These are serious questions which should be answered for all studies which base their findings on the phonetic transcription of spoken data.

Since current sociolinguistic and dialectological research often utilize phonetic transcriptions as the basis for determining (socio-) linguistic shifts and the like, it is crucial that we explore and understand the limitations of the tools we employ for deriving our data sets, especially at the base level of data collection. We must first ascertain that the data, upon which we rely heavily, is in fact valid and not a product of a transcriber's phonetic-phonological idiosyncrasies (recall, for example, Trudgill 1983:38, and problems found with respect to *Survey of English Dialects).*

[1] This research is funded by National Science Foundation Doctoral Dissertation Research Improvement Grant SBR-9313170. I would like to thank my co-authors and H. Paul Manning for contributing in different but important ways to the assembly of the data; in addition I wish to thank Michael Silverstein for his ongoing guidance, insight and suggestions. Any errors or shortcomings are mine alone (L-AL).

In order to address these important issues (cf. Cucchiarini 1993; Nettelbladt 1993; Kerswill & Wright 1990; Vieregge 1987 & 1989; Wright 1983; among others) the validity and reliability of phonetic transcriptions by three graduate linguistic students is being studied. An unfamiliar language was chosen in order to control for native speaker content based filtering of the signal as well as to control for the transcribers being "...influenced by their (phonological) knowledge of the language variety being transcribed." (Kerswill & Wright 1990:258) While a possible argument against the use of non-native transcribers is that the transcribers may not be able to 'hear' the non-native sounds which are to be transcribed, it is not an argument which has been proven in the literature, as for example in the experiments reported on in Vieregge (1987); Wright (1983); Van Valin (1976); Stevens, Liberman, Studdert-Kennedy and Öhman (1969); among others.

2 Methodology

2.1 Data Collection

Based on experiments such as those mentioned above, this study has examined the question of inter- and intra-transcriber reliability as well as validity across two speech styles for a language which the transcribers were previously unfamiliar with. The language being transcribed is a previously undocumented West Jutlandic dialect of Danish, Thyborønsk.

The methodology for this experiment involved briefing the three natively monolingual American English speaking transcribers on minor adjustments made to the IPA from Jespersen's *Danias Lydskrift;* then listening to a set of standardized pronunciation tapes for 10 hours. They were then given two tapes, 10 hours apart. Each of these Style 1 tapes contained a word list (WL) read by an elderly female informant recorded in her home in Thyborøn, Denmark. After approximately 20 hours of training with the Style 1 tapes, a third tape, Style 2, was given to the transcribers, containing utterances (UT) taken sequentially from a longer segment of spontaneous speech from a sociolinguistic interview of an elderly male, recorded at his home in Thyborøn, Denmark. The three test tapes were created in the university's language laboratory by first digitally recording the word lists and phrases from the original analog tapes, then simultaneously recording back onto a digital and an analog tape. Each utterance was recorded onto the test tapes twice with a four second pause between them.

The transcribers were instructed to produce at least two separate transcriptions of each of the tapes. The first pass (conducted at roughly the 20 hour mark for Style 1 and the 40 hour mark for Style 2) was to be done without prior listening or consultation with the master transcription which was produced by a linguist who is a native speaker of the dialect. This was Phase I.1 of the study. The second pass was to be done after the allotted training time was nearly past (at roughly the 50-60 hour mark) and after the master transcription had been consulted. This was Phase I.2 of the study. The transcribers were then each given seven excerpted sociolinguistic interview tapes to transcribe. Upon completion of the 7th interview tape (at roughly the 110 hour mark of transcribing) the transcribers listened to and produced one transcription each of the second of the Style 1 tapes and the Style 2 tape without consultation with the master transcriptions. This was Phase 2 of the study. Phase 3 involved the same methodology as in Phase 2, in that after the 13th interview tape (at roughly the 160 hour mark of transcription) the transcribers again produced a transcription of the two aforementioned tapes. By having the transcribers listen to and produce transcriptions for the same tapes (word lists (WL) and utterances

(UT)) at four controlled stages during the academic year (Phases I.1-III), we have been able to measure their transcriptions across two speech styles (Style 1 and 2).

2.2 Phonological Segment Choice and Testing Methodology

The segment chosen for study was Danish /r/ as it appears in five environments. /r/ was chosen because it generally has a consonantal quality in pre-vocalic environments as in [kro] 'kro', "inn"; and a semi-vowel or vocalic quality in post-vocalic environments, as in [hɔ̈ɤɔ] 'hører', "hear". Furthermore, /r/ has a lowering effect on adjacent vowels increasing the difficulty of producing valid and reliable transcriptions for /r/'s environments. The compounded difficulty of recognizing not only the varying quality of /r/, but also its effect on the immediate environment, makes this an interesting segment to examine for purposes of transcription studies, as well as phonetically and phonologically.

The validity tests involve comparing the transcriptions to the master transcription. If the allophone of /r/ as well as the adjacent segments were transcribed the same as in the master transcription, a point was given. No partial points were assigned. The reliability tests involve both inter- and intra-transcriber tests. Firstly, for those tokens judged valid, an inter-transcriber reliability test was conducted. If 2 or 3 transcribers had the same transcription for the token and its adjacent sounds, and it was a valid transcription, a point was given to each transcriber.

As discussed in Vieregge (1987; 1989), a transcription may be reliable but not valid. Therefore, an inter-transcriber reliability test was conducted for those tokens which did not agree with the master transcription. If 2 or 3 transcribers had the same transcription for the token and its environment, and it was not a valid transcription, a point was given to the transcribers. Finally, an intra-transcriber reliability test was conducted which did not take into account the validity judgment of the token. In these cases a single transcriber's various passes of each style tape were studied for tokens and their adjacent sounds, which were transcribed exactly the same on all passes, and points were assigned accordingly.

3 Original Hypothesis

The original operating hypothesis for this study was:

> *Hypothesis 1*
> Over time and with increased exposure, a transcriber, who is not formally familiar with the language being transcribed, will develop "analytic" (sound perception driven) listening skills and will be able to reach criterion.

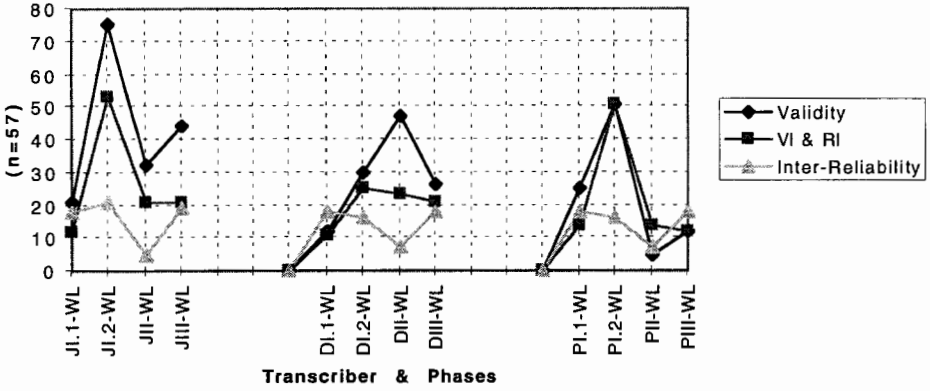
3.1 Validity and Reliability Test Results without Environmental Considerations

To sum up the results from our initial investigation which concerned Phases I.1 and I.2 (as presented in a previous publication), the data revealed that even after the relatively short training time the scores for both validity and reliability for the transcription tests increased, thereby seemingly supporting the original hypothesis.

After the addition of Phase II and III data, a much different picture of the transcribers' scores for validity and reliability tests emerged. Tables A.1 and A.2 represent

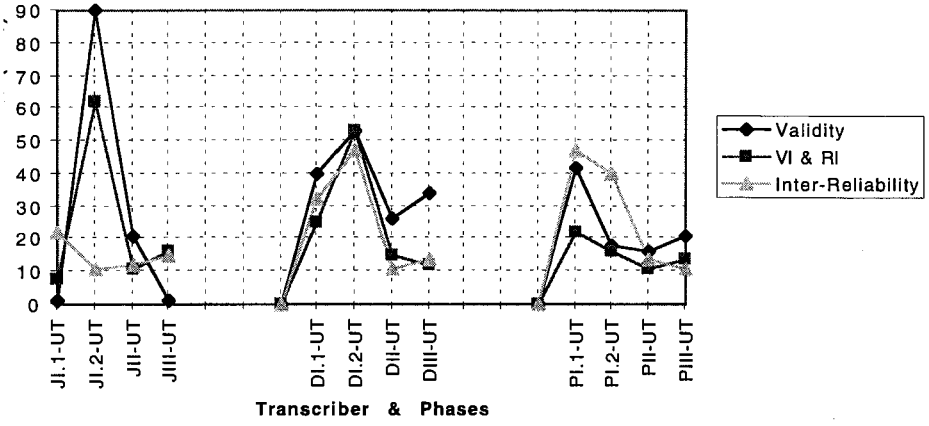
the comparison of the 3 transcribers' scores for the tests as a percentage. Perhaps the most striking *apparent* conclusion based on these graphs is that which could be called a 'learning and unlearning' tendency. Each Table contains three sets of curves. Each set of curves represents the results of the validity, reliability and inter-transcriber reliability test scores for each of the three transcribers. Along the x-axis is listed the transcriber (J, D or P) followed by the Phase number (I.1 through III) and followed by the notation WL or UT (Word List or Utterances) which indicates Style 1 or 2. To the far right of the graphs is a key box wherein information can be found as to how each test is graphically represented. The y-axis represents the score in percent for the number of transcribed tokens which correspond to the test being considered, in parenthesis is offered the number of tokens in the sample.

Table A.1: Phases I-III: Style 1: Comparison of Transcribers' Scores on Validity and Reliability Tests in %



If we consider the direction of the majority of the curves for both Table A.1 and A.2, we find that during Phases I.1 and I.2, the general tendency is an upward direction, this is what we refer to as the expected 'learning' period. However, Phase II, for the most part, often reveals a dramatic decrease in test scores followed by a slight increase for Phase III. This could suggest that perhaps the transcribers initially learn to hear the /r/ sounds and their environments at a dramatic rate, but after increased exposure, they somehow 'unlearn' the initial sound categorization which they had developed by the end of Phase I.2. By the end of Phase III (now roughly 160 hours into interview transcriptions), it appears that they have reconstructed a working phonology of the language which generally shows a moderate success rate, revealed by the upward movement from Phase II to III. Of course, there are individual test and transcriber differences, but the strikingly similar pattern for the majority of these curves is as just described. This should remind us of Bloomfield's rule of thumb, always throw away the first three months of your transcriptions.

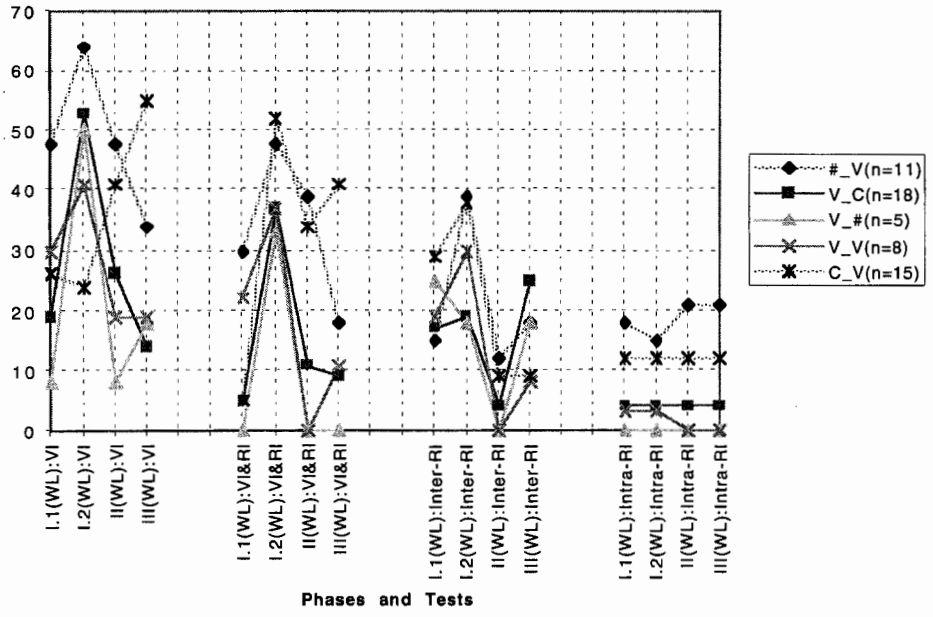
Table A.2: Phases I-III: Style 2: Comparison of Transcribers' Scores on Validity and Reliability Tests in %



3.2 Validity and Reliability Test Results with Environmental Considerations

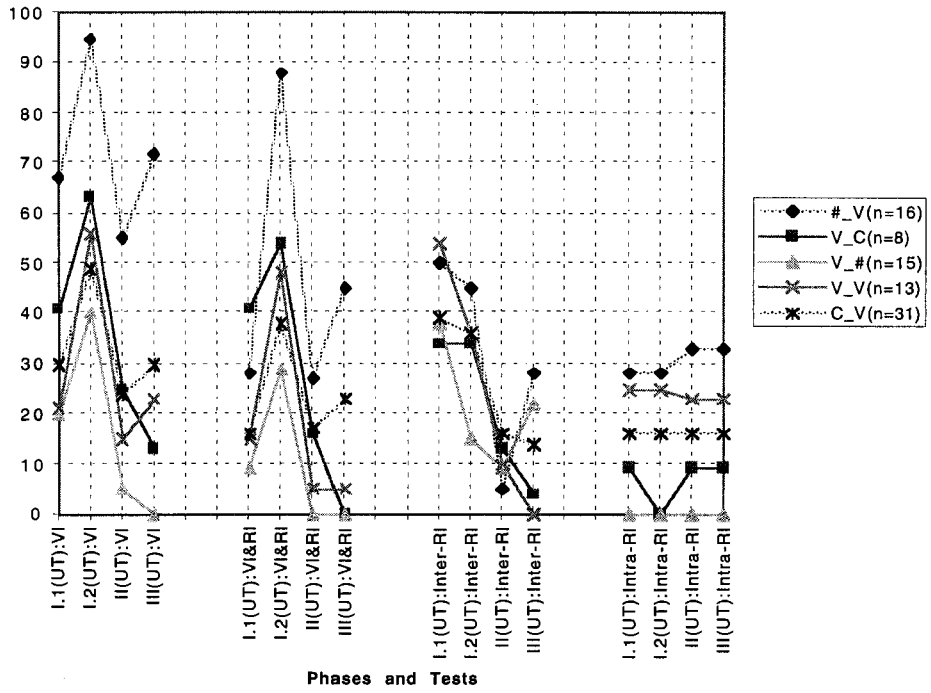
In order to better understand what may be causing the trends shown in Tables A.1 and A.2, an examination was conducted of /r/'s environments as a function of the validity and reliability tests. The results of which are represented in Tables B.1 and B.2, wherein the average score (as a percent) on the validity and reliability tests for all transcriptions of /r/ are plotted by environment across time for both Styles 1 and 2, respectively.

Table B.1: Average Score (in %) on Validity and Reliability Tests for All Transcriptions of /r/ in Style 1 by Environment



Again we immediately notice the general up-down-up tendency for the validity and reliability tests. Note the pattern of Phase I.1-I.2 upward (or increasing score) direction of the curves, followed by a downward (or decreasing score) direction of the curves between Phases I.2-II, with a slight resurgence in scores, shown by the upward direction of the curve from Phase II-III. This tells us that the shapes of the curves in Tables A were not a result of simply having added the environments together, as we find similar curves in Tables B. Rather, our conclusion must be that there exists another variable at play which is influencing the graph shapes in a consistent and predictable manner. The intra-transcriber reliability curves expectedly show a different patterning, as they reflect the average score for an individual transcriber's degree of reliability (or consistency) across his or her own transcriptions. The intra-transcriber reliability curve is more or less a straight line, revealing the percent level for which a given transcription is produced exactly the same by all transcribers in the environment represented by that line.

Table B.2: Average Score (in %) on Validity and Reliability Tests for All Transcriptions of /r/ in Style 2 by Environment



3.3 Rank Ordering of Environments

Table B.3 presents the rank ordering of the environments for /r/ from highest to lowest scoring for each test, as well as the overall ranking of the environments for all tests. This data is taken directly from Tables B.1 and B.2. The result is an interesting pattern which immediately clues us in to an important variable which has more than likely affected the data viewed thus far.

	Validity:	Validity & Reliability:	Intertranscriber Reliability:	Intratranscriber Reliability:		**Overall Ranking:**
Style 1 (WL)	#_V	#_V	#_V	#_V		**#_V**
	C_V	C_V	C_V	C_V		**C_V**
	V_V	V_V	V_#	V_C		**V_V**
	V_C	V_C	V_C	V_V		**V_C**
	V_#	V_#	V_V	V_#		**V_#**
Style 2 (UT)	#_V	#_V	#_V	#_V		**#_V**
	V_C	V_C	C_V	V_V		**C_V**
	C_V	C_V	V_V	C_V		**V_C**
	V_V	V_V	V_C	V_C		**V_V**
	V_#	V_#	V_#	V_#		**V_#**

Table B.3: Rank ordering of the environments for /r/ from highest to lowest score for each test

From Table B.3, our reading of the earlier graphs is confirmed. The ranking of environments is fairly consistent, not only across tests, but also across style. Namely, that pre-vocalic /r/ environments are highest ranked for test score percentage while post-vocalic /r/ environments are lowest ranked. The fact that the rank ordering is so consistent across tests and styles forces a reconsideration of the impact which /r/'s different allophones may have on the transcriptions. Factors to be considered are: (1) the allophonic distribution of /r/; (2) the differing phonetic quality of the /r/ allophones (i.e., consonantal and vocalic); and (3) the comparative functional load of /r/ in the native language of the transcribers and in the transcribed language (all three transcribers are native speakers of only Standard American English).[2]

As discussed earlier, Danish /r/ has two allophones which are phonetically quite distinct. The allophones are in complementary distribution in pre-vocalic positions, where only [r] occurs, as in [tres] 'tres', "thirty", and in post-vocalic positions, where only [ɹ] occurs, as in [fʏ̈ɹə] 'fyrre', "forty". However, the two allophones are in free variation in intervocalic positions, as in [æ.ɹøʔ] 'Ærø' (name of a Danish island) and [fæ.ɹøʔɔnə] 'Færøerne', "Faeroese Islands" (cf. Heger 1981:30-31). Table B.4 represents the distribution of Danish /r/. [r] is the consonantal allophone, and is the more distant of the two allophones to the phonology of American English, as we find a close relative to Danish [ɹ] in American English.

[2] It must be remembered that some dialects of American English do have distinct allophones of /r/ and Standard American English will most likely have a "stronger" prevocalic [r] than a postvocalic one. In other words, Standard American English could be considered to have two or three main allophones of /r/: unstressed (probably syllabic) [ɚ], prevocalic approximate [ɹ], and a weaker postvocalic articulation of approximant [ɹ̞]. Danish [ɹ̞] might be indistinguishable from the [ɹ] of many speakers of Standard American English. At this point tests on the speech of the transcribers have not been conducted. This is, however, an area which is planned for study in association with the ongoing experiments in order to shed light on issues such as functional load differences and similarities as they impact the transcription process.

	[r]		[ɹ]
Environments of Complementary Distribution:	#_V C_V		V_# V_C
Environment of Free Variation:		V_V	

Table B.4: The distribution of Danish /r/

Tables B.1 through B.4 underscore the probability that the transcription scores are being influenced by some or all of the factors relating to the allophones of /r/, as just outlined. Therefore, we must reconsider our original hypothesis to include and highlight this variable in further tests.

4 Re-examination of /r/ and phonological categorization

Based on the analysis of the data and the consideration of additional variables, as presented above in Section 2, the revised hypothesis reads as follows: Learning to produce phonetic transcriptions is a function of structural push and pull in which the allophone level is relevant to understanding the transcriber's ability to reach criterion.

4.1 Methodology

In order to eliminate the additional variable of possible errors in the native speakers' transcriptions of the word list and phrases, a new master transcription was created, this time based on the actual sounds produced by the informants as well as abiding by the phonological rules for the distribution of /r/. Despite the (Danish) native speaker linguist being well-trained in transcription methods, the influence of orthographic knowledge on the transcription of inter-vocalic /r/ positions was quite striking. Unfortunately, this research question must be reserved for future work. The scoring of points for transcribing either /r/ allophone no longer included a consideration of the adjacent sounds, as was the case for the validity and reliability tests. All transcriptions of [r] and [ɹ] were counted and scored according to the environment in which the transcriber placed the token.

The reason for the departure from validity and reliability tests as outlined in Sections 1 and 2 is based on problems encountered with establishing empirical means for determining validity and reliability in transcriptions. In the case of validity testing, we were relying on the comparison of non-native to native speaker produced transcriptions. As mentioned above, close examination of the native speaker linguist's transcriptions showed notable influence from orthographic knowledge, thereby calling into question the basis on which validity judgments were made. Similarly there exists fundamental problems with the notion of reliability, as Cucchiarini (1993) elegantly points out:

> While agreement indicates to what extent a number of objects are given identical ratings by different subjects, reliability reflects the degree to which the relationships between the different objects are judged in the same way by the subjects. ...To sum up, agreement concerns the absolute values of the ratings, whereas reliability represents to what extent they vary in the same way or, put otherwise, 'the degree to which the ratings of different

> judges are proportional when expressed as deviations from their means' (Tinsley and Weiss 1975:359). ...Since the definition of reliability is based on the notion of proportionality, determining reliability presupposes at least an interval level of measurement. ...Given that observations about transcriptions are not amenable to interpretation in terms of mean, deviation from the mean, and variance, reliability cannot be calculated. (1993: 65-66)

As a result of concerns with the use of terms such as 'validity' and 'reliability', we have opted to explore the evaluation of transcriptions in terms of 'reaching criterion'. The intent of this terminology, 'reaching criterion', is to indicate whether a linguist (or a group of linguists) can attain a level of transcription which would be accepted as representational of the spoken data. The evaluation of such attainment is, in turn, based on knowledge of the systematicity of the speech norm as well as on the agreement of the transcription to the spoken data. As Cucchiarini states:

> Given that there are no such things as THE feature set of THE feature hierarchy for transcription evaluation, it seems that any decision will have to be based on a number of assumption [sic] that have to be reckoned with in evaluating the results obtained. This may imply that agreement between transcriptions cannot be established in absolute terms, but has to be related to specific research goals. (1993:88)

4.2 Testing and results

The scores for all three transcribers were averaged and then converted into percentages based on the total possible number of /r/'s as recorded in the new master transcription. The total n for each style is given along the y-axis of each graph, and the individual n's for each of the two /r/ allophones, as recorded from the master transcription, are given in the log boxes to the right of the graph.[3] In considering Tables C.1 and C.2, we must keep in mind the allophonic distribution of /r/ as laid out in Table B.4. Namely, that phonologically only [r] appears in the two pre-vocalic environment, and that phonologically only [ɹ] appears in the two post-vocalic environment, while in inter-vocalic position, [r] and [ɹ] are in free variation. Therefore, in considering the graphs for each allophone, we would expect that the two environments which favor the allophone will show the highest score percentages, conversely, the two environments which disfavor the allophone will show the lowest score percentages. In the case of the inter-vocalic environment, we would expect to find a median scoring tendency as either allophone may occur. The fact that /r/ does not have consonantal and vocalic reflexes with a complex allophonic distribution in Standard American English, as it does in Danish, must also be considered (recall footnote 2).

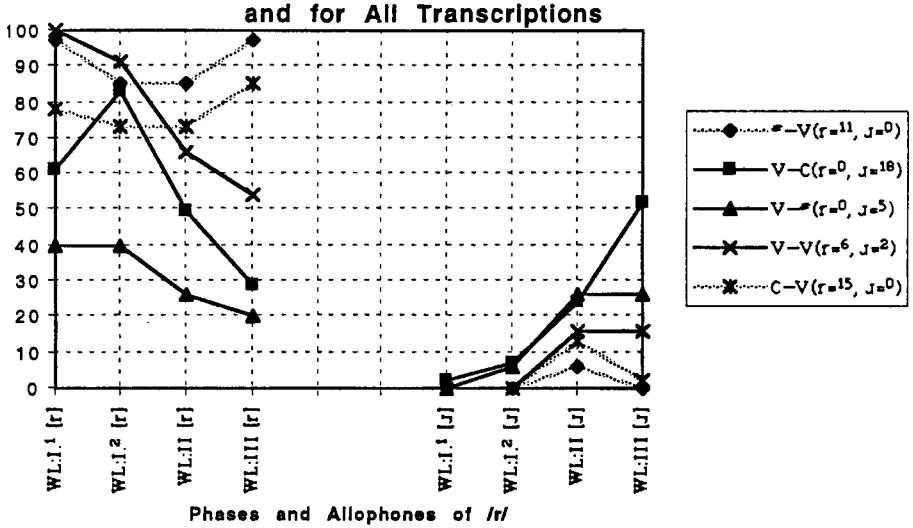
Furthermore, the functional load of [r] is zero in American English, while the American sister to [ɹ],[ɾ]/[ɹ], does carry a functional load in the transcribers' native language. This might lead us to predict that the transcribers will be more likely to use [ɹ] more often, hence have higher scores than for [r], as it is a close relative to their native [ɾ]/[ɹ] sound and they are more accustomed to hearing or listening to that sound. On the

[3] The determination of total n and individual n for each of the two /r/ allophones is still problematic since the population counts are still based on the master transcription. However, the master transcription used to produce the population counts was first revised such that orthographic influence on the transcription was removed. Presently under consideration are various means for improving the counting the sample populations, in order to achieve a more empirical basis for determining such data.

other hand, we might predict that despite the closeness of the sound of [ɹ] to the American /r/, the consonantal quality of the [r] allophone is less likely to not be heard or to be mistaken for another sound, therefore more likely to be identified and transcribed.

Tables C.1 and C.2 represent the average percentage of tokens of each /r/ allophone transcribed in each of the five environments as a function of time.

Table C.1: Average % of Use of /r/'s Two Allophones (Style 1): Viewed by Environment, Across All Phases and for All Transcriptions

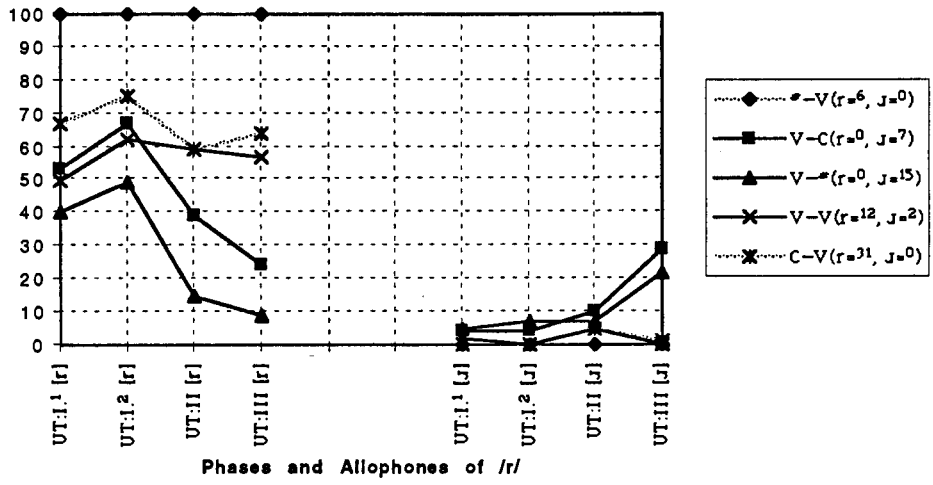


As we see in C.1 for the [r] allophone, the two pre-vocalic curves are parallel and show an overall tendency of increasing scores, with a slight decrease from Phase I.1 to I.2. The expected result is that these two curves would be higher in that pre-vocalic environments are phonologically predicted for [r]. Again as expected, the intervocalic curve lies in the middle of the five curves, as it is the environment of free variation for /r/. The two post-vocalic environments are the lowest curves for [r], despite a sudden peak at Phase I.2 for the post-vocalic pre-consonantal position. Again, phonologically we do not predict any occurrences of [r] in a post-vocalic environment.

While the curves for the [ɹ] allophone are much more condensed, there is still evidence for the phonologically predicted post-vocalic environments to be favored over the pre-vocalic environments. Common to both allophones is that by Phase II the predicted environment scores are unquestionably higher than the environments in which they were

not predicted to occur. Furthermore, for both allophones we find that the intervocalic environments maintain a mid range position.

Table C.2: Average % of Use of /r/'s Two Allophones (Style 2): Viewed by Environment, Across All Phases and for All Transcriptions



Similar to the reading of Table C.1, Table C.2 supports the predicted pattern. Namely, the phonologically predicted environments for each allophone are favored. The data for the inter-vocalic environment, in which the allophones of /r/ are in free-variation, center around a mid-range position.

One means of evaluating the results from these two tables, is to once again rank order the environments by most to least frequently used according to each /r/ allophone. The results of rank ordering are presented in Table C.3. These results confirm our suspicions that the transcribers are gradually conforming to the allophonic distribution of /r/, despite having no formal information about the phonology of the dialect, and only being exposed to it through the transcribing tapes. We find that the overall ranking of [r]'s environments shows the pre-vocalic environments ranked highest, inter-vocalic is ranked in the middle, and the post-vocalic positions are ranked lowest. As we would expect, the opposite rank ordering for pre- and post-vocalic environments is found for [ɹ].

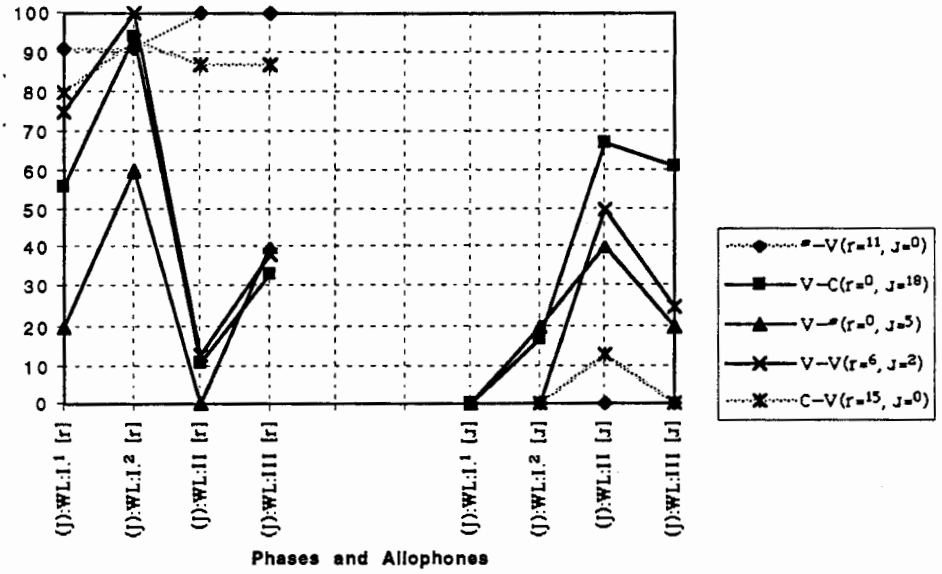
	[r]	[ɪ]		Overall Ranking for [r]:	Overall Ranking for [ɪ]:
Style 1 (WL):	#_V	V_C			
	V_V	V_#			
	C_V	V_V			
	V_C	C_V		#_V	V_C
	V_#	#_V		C_V	V_#
				V_V	V_V
				V_C	C_V
Style 2 (UT):	#_V	V_C/V_#		V_#	#_V
	C_V				
	V_V	V_V			
	V_C	C_V			
	V_#	#_V			

Table C.3: Overall rank orderings of environments for Danish /r/ allophones

4.3 Individual transcriber results

The D tables show the comparison (in percentages) of the use of /r/'s two allophones as viewed by environment and across time. This section is graphically represented in six tables, D.1.1 through D.2.3. D.1 refers to Style 1 (WL) and D.2 refers to Style 2 (UT) while the final number, ranging from 1 to 3, refers to the individual transcriber whose scores are being considered. Along the x-axis the transcriber is represented in parenthesis by a letter (J, D or P), the style code follows (either WL or UT), thereafter appears the phase number (from I.1 through III), and finally the allophone under consideration ([r] or [ɪ]).

Table D.1.1: Phases I-III: Style 1: J: Comparison of Use of /r/'s Two Allophones: Viewed by Environment and Across All Phases

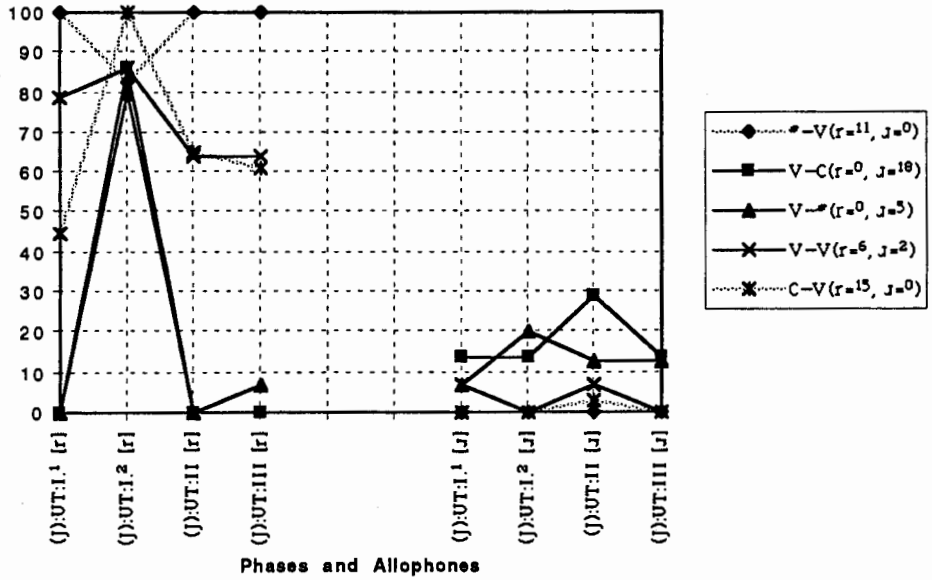


If we consider J's curves (Tables D.1.1 and D.2.1), we notice the following: In Table D.1.1 for the [r] allophone, the pre-vocalic environments are consistently highly ranked. This is what we expected given that they are the phonologically predicted environments. The graphs also reveal a lower rate of fluctuation in score over time and the scores plateau by Phases II and III. The post-vocalic positions for [r] score lower than the predicted pre-vocalic positions, again, as expected.

Interestingly enough, we notice a re-emergence of the up-down-up curve. A possible explanation for these curves is an initial favoring of the [r] allophone, extending it to all cases of /r/. The inter-vocalic environment is initially located in a mid range position according to token scoring, but then during this transcriber's period of [r] abundance, it shoots to being used 100% of the time in this environment. Since the inter-vocalic position is unpredictable for allophone occurrence, we would expect to find, as indeed we do, different approaches from each of the transcribers to solving the distribution of /r/ when its allophones are in free-variation. By Phases II and III, the inter-vocalic position has dropped and continues to parallel the post-vocalic environments' curves.

In considering J's [ɹ] curves, we immediately note the phonologically predicted low scores for the pre-vocalic environments. The post-vocalic positions show a general learning curve with a fall off by Phase III. This too is to be expected if we look in the log box and note the infrequency of [ɹ]in the environments. This is unfortunately an artifact of the size of the current data set.

Table D.2.1: Phases I-III: Style 2: J: Comparison of /r/'s Two Allophones: Viewed by Environment and Across All Phases



For J, there does seem to be a difference in scoring related to style. This is evidenced by the difference between Tables D.1.1 and D.2.1. In D.2.1, we are immediately struck by the different graph shapes which highlight the difference in strategies taken in transcribing the utterances as opposed to the word list. If we examine the graphs more closely, we find that there continue to be some important similarities between the two speech style transcriptions. Namely, we see that for the [r] allophone, the phonologically predicted pre-vocalic environments are again ranked highest as an overall trend. However, we cannot disregard the fact that in C_V environments J's scores parallel the post-vocalic curves while for Style 1, J's C_V scores paralleled the #_V scores. What we may conclude from this is at the level of conjecture, as we would require further data from an additional Style tape to

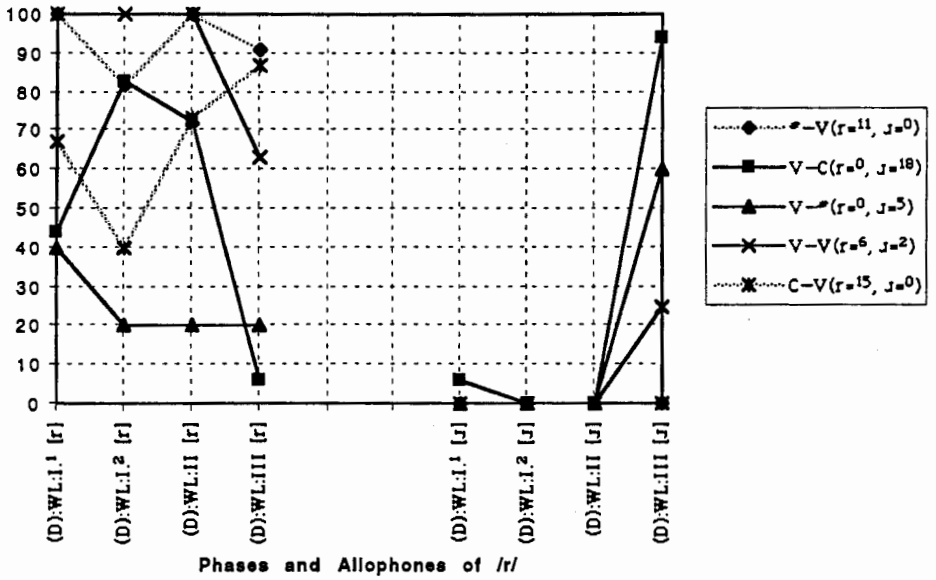
draw any conclusions. This type of data is currently in the process of being coded for such a study. Of additional importance, we recall that the post-vocalic curves for [r] are very similar for both Style 1 and 2. This similarity points to the possibility of a pattern in the data. Namely that in transcribing, or in learning to transcribe, the [r] allophone is not likely to be identified in post-vocalic environments by the three transcribers studied. This is reflected in the dramatic drop rates from Phase I.2 to II. The pattern is noticeably stronger for Style 2 than for Style 1.

In considering [ɪ], we again see that there are surface differences between Tables D.1.1 and D.2.1. It is important to note the re-emergence of the tendency of [ɪ] being transcribed in the post-vocalic environments more frequently than in pre-vocalic environments. Pre-vocalic [ɪ] is generally only transcribed between 0-10% of the time; this is in line with phonological predictions for the allophonic distribution of /r/. Again, as in D.1.1, the scoring for inter-vocalic environment is roughly located between the two pre- and two post-vocalic curves.

Table D.1.2 represents D's scores for the Style 1 tape. It is here we first encounter an example of the difference in individual strategies towards solving the complex distribution of /r/ in Danish as briefly mentioned above. D's graphs of scores for [r] are quite different from J's. We note that initially D strongly favors not only the predicted #_V environment, but also the inter-vocalic and the V_C environments. While one might attempt to rationalize this behavior for the inter-vocalic environment by recalling that it is an area of free variation, hence a phonologically sanctioned environment for [r], we cannot employ such an explanation for the V_C environment. Therefore, it is suggested that the best means for understanding these curves, and Tables D.1.1 through D.2.3 in general, is by noting the individual strategies and keeping in mind the question of whether or not the transcribers reach a stage (and at what point) where their distribution of /r/ matches the phonologically predicted distribution. For this question the answer is yes, the transcribers do reach a stage (generally at Phase II or III) where their distribution for /r/ does reach criterion. We note that while some environments of [r] for this transcriber are immediately sorted out and are in phonological 'agreement' (note the low scores for V_# and high scores for #_V), other environments take a longer time to be sorted out (note the inter-vocalic and C_V curves) but by Phase III, the environment ranking is in line with what we would predict.

D's [ɪ] scores pattern in a considerably more obvious way than D's [r] scores. Quite simply stated, D does not display any use of [ɪ] until Phase III, where the distribution is in perfect alignment with the predicted patterning of post-vocalic highest, inter-vocalic mid level, and pre-vocalic lowest (in this case at 0% use).

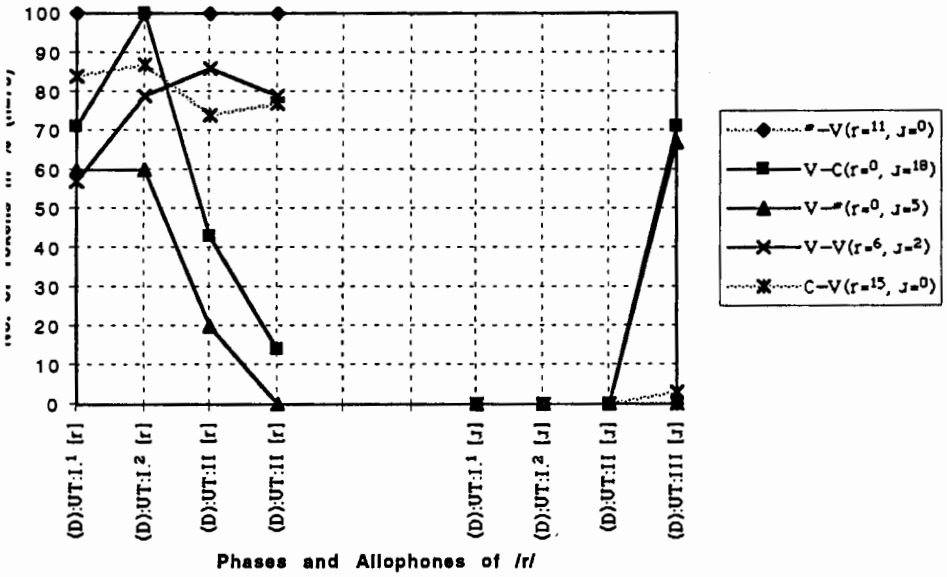
Table D.1.2: Phases I-III: Style 1:D: Comparison of Use of /r/'s Allophones: Viewed by Environment and Across All Phases



The results from this transcriber's Style 2 data show a more systematic strategy for resolving the distribution of [r] and reaching criterion. The use of [r] in word initial pre-vocalic position is consistently ranked highest, with 100% of occurrences being transcribed as predicted, thereafter the post-consonantal pre-vocalic environment followed by the inter-vocalic environment. The Phase III score of 0% use of [r] in word final post-vocalic position is equally important to criterion judgments as the word initial 100% ranking.

Unquestionably, the Phase III scores also reveal a clear distribution for the [ɹ] allophone. Specifically we note that the phonologically predicted post-vocalic environments rank highest (at roughly the 70% mark). It is clear from Table D.2.2 that by Phase III there is strong evidence supporting transcriber D as having resolved the phonological distribution of /r/.

Table D.2.2: Phases I-III: Style 2: D: Comparison of Use of /r/'s Two Allophones: Viewed by Environment and Across All Phases



Based on Tables D.1.3 and D.2.3, it is clear that by Phase III the third transcriber's scores reflect the predicted distribution of /r/. Considering the scores for [r] in Table D.1.3, we note that for Phases I.1 and I.2, the expected distribution based on environment is evidenced. The Phase III curve is somewhat difficult to interpret, other than it most likely being P's point of 'unlearning' before final phonological categorization which is seen in the Phase III scores. Having a period of 'unlearning' is both reminiscent of the other transcribers' curves (though this point occurs at different Phases and/or different Styles for any given transcriber) as well as the initial A and B Tables where we first noted the 'up-down-up' or 'learning-unlearning' curve shapes. This finding, that individual transcribers have different strategies for developing a categorization for the allophonic distribution of Danish /r/, contributes not only to the further understanding of the learning process of individual transcribers but also speaks to questions raised in the interpretation of more generalized graphs such as those in A and B.

The [ɹ] curves are more difficult to interpret for this transcriber, as we not only find the predicted post-vocalic environments among the highest ranked, but also the C_V environment. Recall that C_V is not phonologically predicted for [ɹ]. Due to the small

sample size, it is only possible to speculate as to the reasons behind this occurrence. Therefore, we will defer discussion of this point until more data can be analyzed, thereby allowing more statistically valid results upon which conclusions may be based.

Table D.1.3: Phases I-III: Style 1: P: Comparison of Use of /r/'s Two Allophones: Viewed by Environment and Across All Phases

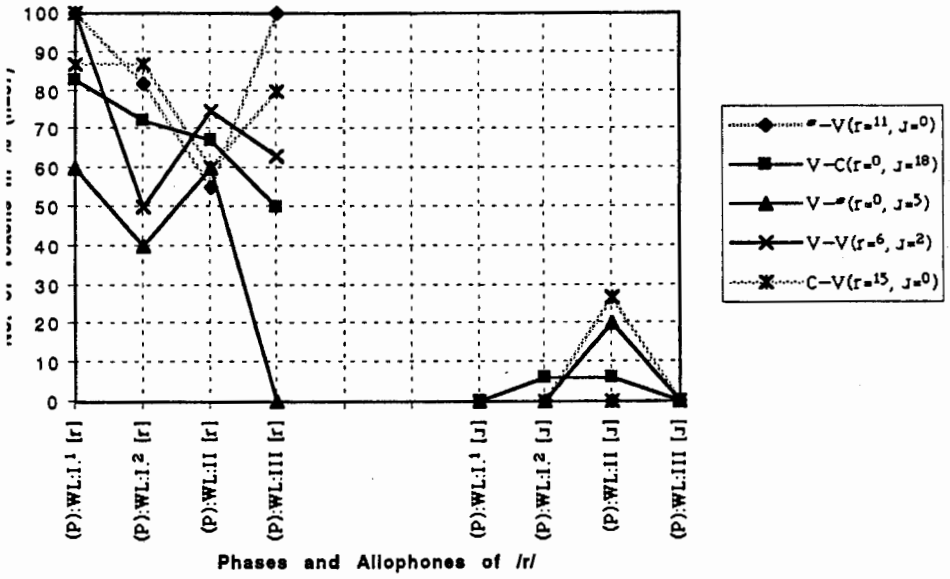
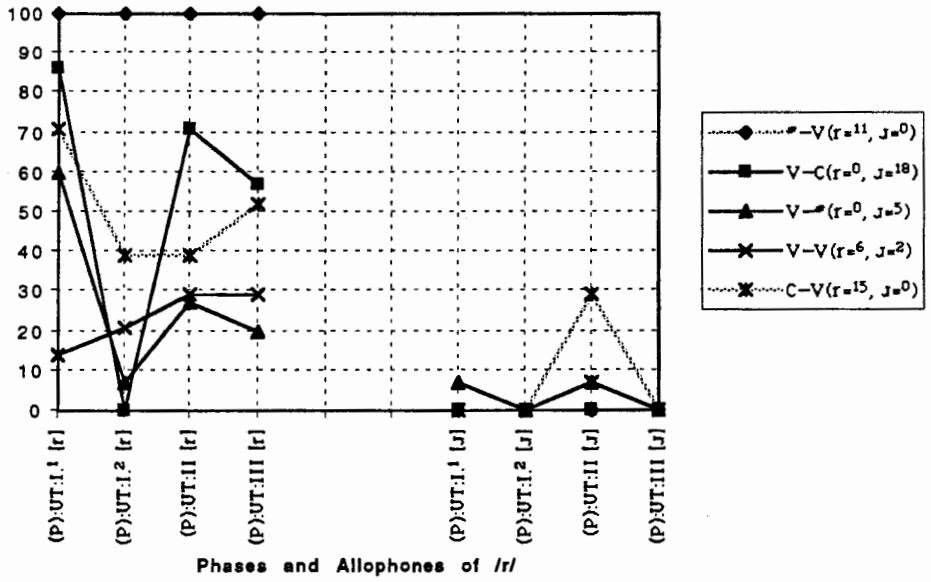


Table D.2.3 presents interesting questions as to P's strategies for conforming to the phonologically predicted allophonic distribution of Danish /r/. Similar to P's Style 1 curves for [r] as well as D's Style 2 curves for [r], in P's Style 2 we note that the word initial pre-vocalic position is consistently ranked at 100% use. While the other environments seem to undergo considerable re-shuffling, the final (i.e. Phase III) distribution matches the predicted distribution for all but one environment, V_C. The curves for [ɹ] pattern similarly to those for P's Style 1 results, leading us to the same conclusion that more data is needed in order to empirically address the findings.

Table D.2.3: Phases I-III: Style 2: P: Comparison of Use of /r/'s Two Allophones: Viewed by Environment and Across All Phases



4.4 Result summary of individual transcribers' tables

An interesting point to be drawn from considering the three transcribers' individual graphs for the distribution of the use of the two /r/ allophones, is that we are immediately struck by how differently each transcriber penetrates the Danish system on an environment by environment trend. We notice that regardless of their individual tactics, we find that by Phase III, the transcribers are generally in conformity with the phonologically predicted allophonic distribution of Danish /r/. Even more striking is their individual overall rankings of the environments by allophone for highest to lowest use, as presented in Table D.3. These results reveal that taken as a whole, the transcribers' environment ranking is *not* sensitive to speech style for the data studied thus far. It is interesting that for [r] the transcribers' score for the inter-vocalic position (i.e. V_V) was higher than the post-consonantal/pre-vocalic position (i.e. C_V), meaning that they more frequently put [r] in an inter-vocalic position than they did in a post-consonantal position. Whether this is purely an artifact of the size of the data set or actually a tendency revealing something about a non-native speaker transcriber's approach to penetrating the allophonic distribution of /r/ when it

appears in free variation has not yet been determined. As for the environmental ranking of [ɹ], we find the scale which we had predicted based on the phonological distribution characteristics.

	[r]	[ɹ]
Table D.1.1: (J):WL	#_V	V_C
	C_V	V_V
	V_V	V_#
	V_C	C_V
	V_#	#_V
Table D.2.1: (J):UT:	#_V	V_C
	V_V	V_#
	C_V	V_V
	V_C	C_V
	V_#	#_V
Table D.1.2: (D):WL	#_V/V_V	V_C
		V_#
	V_C	V_V
	C_V	C_V/#_V
	V_#	
Table D.2.2: (D):UT:	#_V	V_C
	V_V	V_#
	C_V	C_V
	V_C	V_V/#_V
	V_#	
Table D.1.3: (P):WL	#_V	C_V
	V_V	V_#
	C_V	V_C
	V_C	V_V/#_V
	V_#	
Table D.2.3: (P):UT:	#_V	C_V/V_#
	V_C	V_V
	C_V	V_C
	V_V	#_V
	V_#	

Overall Ranking for [r]:Style 1:	Overall Ranking for [ɹ]:Style 1:
#_V	V_C
V_V	V_#
C_V	V_V
V_C	C_V
V_#	#_V

Overall Ranking for [r]:Style 2:	Overall Ranking for [ɹ]:Style 2:
#_V	V_C
V_V	V_#
C_V	V_V
V_C	C_V
V_#	#_V

Table D.3 Individual and overall rank orderings of environments for Danish /r/ allophones

5 Conclusion

In sum, the data support the revised hypothesis.

> *Hypothesis 1'*
> Learning to produce phonetic transcriptions is a function of structural push and pull in which the allophone level is relevant to understanding the transcriber's ability to reach criterion.

This was demonstrated by the various graphs represented in Tables C and D. The D tables provided more detailed information as the micro-diachrony of how the individual transcribers worked through the two phonetic variants of Danish /r/ and were finally able to penetrate the Danish system, and begin to reach criterion in their transcriptions for these two allophones.

In conclusion, it is worthwhile to underscore that if we are indeed trying to understand and determine the existence of various socio- and/or ethnolinguistic shifts and the like, we must honestly evaluate and understand the limitations of the tools which we employ for deriving our data sets, especially at the base level of data collection. Without such understanding, we may run the risk of blindly distorting our models and theories.

6 Directions for Future Research

There are numerous questions which are raised by this work. Presently we are examining other phonemes to contrast with /r/'s distribution. The goal is to determine whether the findings presented herein are unique to /r/ or if, as we suspect, they do indeed speak to the larger question of the evaluation and further understanding of both the processes undergone by transcribers in transcribing and the final transcriptions produced. In addition to other phonemes, another speech style (excerpted interviews) is being studied to collect additional data both to increase the total population sampled as well as to address questions of style in transcriptions as raised in Section 3.3. It will also be possible to monitor at least one of the transcribers over additional phases, allowing for continued mapping of learning curves, strategy development (i.e. allophonic distribution) and increases or decreases in reaching criterion in transcriptions. It is hoped that through further study we may better address the question of whether or not it is possible for linguists to reach criterion in phonetic transcriptions, and the question of how we are to evaluate the transcriptions which we utilize on a regular basis.

References

Cucchiarini, C. (1993). *Phonetic transcription: a methodological and empirical study* (Nijmegen: Katholieke Universiteit Nijmegen).

Dressler, Wolfgang U. and Wodak, Ruth (1982). "Sociophonological methods in the study of sociolinguistic variation in Viennese German", *Language in Society.* 11:339-370.

Greenberg, J.H. (1967). "The first (and perhaps only) non-linguistic distinctive feature analysis", *Word* 23:214-220.

Greenberg, J.H. and Jenkins, J.J. (1964). "Studies in the psychological correlates of the sound system of American English, I and II", *Word* 2:157-178.

Heger, S. (1981). *Sprog og Lyd, elementær dansk fonetik* (København: Akademisk Forlag).

Hockett, C.F. (1967). "The quantification of functional load. Word. 23:300-320.

Jespersen, O. (1890-1892). "Danias lydskrift", *Dania; tidsskrift for folkemål og folkeminder* 1:33-79.

Kerswill, P.E. (1985). "A sociophonetic study of connected speech processes in Cambridge English: An outline and some results", *Cambridge Papers in Phonetics and Experimental Linguistics* 4.

Kerswill, Paul and Wright, S. (1990). "The validity of phonetic transcription: Limitations of a sociolinguistic research tool", *Language Variation and Change* 2, 3:255-276.

Ladefoged, P. (1967). *Three Areas of Experimental Phonetics* (London: Oxford University Press).

Ladefoged, P. (1982). *A Course in Phonetics* (second edition) (Blackpool: Harcourt, Brace, Jovanovich).

Lane, Lisa Ann, Denton, J. & Suslak, D. (1995 forthcoming). "The validity and reliability of phonetic transcriptions for sociolinguistics and dialectology" *SALSA III. Proceedings of the Third Annual Symposium about Language and Society - Austin.* (Austin: The University of Texas, Department of Linguistics, Austin, Texas).

Miller, G. A. (1963). *Language and Communication* (New York: McGraw-Hill Book Co.), 10-79.

Nettelbladt, U. (1993). "Some reflections on transcribing", *Gothenburg Papers in Theoretical Linguistics 72.* (Proceedings of the XIVth Scandinavian Conference of Linguistics and the VIIIth Conference of Nordic and General Linguistics, August 16-21, 1993), 47-62.

Stevens, K.N., Liberman, A.M., Studdert-Kennedy, M. and Öhman, S.E.G. (1969). "Crosslanguage study of vowel perception", *Language and Speech* 12:1-23.

Tinsley, H.E.A. and D.J. Weiss. (1975). "Interrater reliability and agreement of subjective judgments", *Journal of Counseling Psychology* 22:358-376.

Trudgill, Peter (1983). *On Dialect* (Oxford: Blackwell).

van Valin Jr., R.D. (1976). "Perceived distance between vowel stimuli", *Journal of Phonetics* 4:51-58.

Vieregge, W.H. (1987). "Basic aspects of phonetic segmental transcription", *Zeitschrift für Dialektologie und Linguistik (Beiheft Nr. 54, Probleme der phonetischen Transkription)* (Wiesbaden: Steiner), 5-55.

Vieregge, W.H. (1989). *Phonetische Transcription Theorie und Praxis der Symbolphonetik.* (Stuttgart: Steiner).

Vieregge, W.H., & Cucchiarini, C. (1989). "Agreement procedures in phonetic segmental transcriptions". In M.E.H. Schouten and P.T. van Reenen, eds., *New methods in dialectology: Proceedings of a workshop held at the Free University, Amsterdam, December 1987* (Dordrecht: Foris), 37-44.

Wright, S. (1983). "The effect of native language on the perceived distance between vowels" *Cambridge Papers in Phonetics and Experimental Linguistics* 2.

Wright, S. (1986). "The interaction of sociolinguistic and phonetically-conditioned CSP's in Cambridge English: Auditory and electropalatographic evidence", *Cambridge Papers in Phonetics and Experimental Linguistics.* 5.