AN EMPIRICAL APPROACH TO THE CODING OF TRANSITIONAL CONTINUITY AND TERMINAL PITCH DIRECTION IN SPOKEN AMERICAN ENGLISH

Robert Englebretson
University of California Santa Barbara

1 Introduction

This paper presents the results of an experiment designed to assess interrater reliability for the coding of two prosodic constructs used in the transcription of spoken language[1] Specifically, I will discuss perceptions by native American-English speakers of (i) the functional category of Transitional Continuity and (ii) the acoustic category of Terminal Pitch Direction, as proposed in the transcription system developed by Du Bois et al (1992, 1993) I will begin with a general overview of this transcription system, followed by a discussion of the experiment and results

While the aim of this paper is to assess interrater reliability in a rather specific area of transcription research, it is hoped that the methodology will be generally applicable to other areas of linguistic research as well As in other social sciences such as experimental psychology or sociology (cf Bakeman and Gottman 1986), linguistic research relies heavily on the coding of observed phenomena The judgement of sentences as "grammatical" or "ungrammatical" by traditional generativists, the recognition and characterization of gestures and interactional strategies by conversation analysts, the classification of child utterance by language acquisition researchers, and the identification of clause types and morphological structures by typologists and descriptive grammarians, all depend to a certain degree on the researcher's perception and interpretation of observed data Different researchers can, and often do, perceive and interpret the same observations in very different ways In many instances, the reliability of a given theory is only as good as the reliability of the researchers who observe the data Thus the notion of interrater reliability is just as crucial for linguistics as it is in other fields, and the quantitative methodology for its assessment which I present in this paper is also applicable to other areas of linguistic research

2 Background

For a detailed description of the Du Bois et al transcription system, consult Du Bois et al (1992, 1993), as I will only outline features relevant to the present paper One of the large-scale uses of the du Bois et al system has been the Santa Barbara Corpus of Spoken American English (SBCSAE) The SBCSAE is an ongoing project at the University of California Santa Barbara, involving the collection and transcription of recordings of naturally-occurring spontaneous spoken

[1] I would like to express my sincere gratitude to the twelve raters who gave of their time, energy, and perceptions I gratefully acknowledge Prof Rebecca Zwick of the UCSB Graduate School of Education, for advice on statistics--any misunderstandings or misapplication of methodology or results are solely my responsibility For access to data, and for invaluable training and experience working with spoken English, I gratefully acknowledge Prof John Du Bois and the Corpus of Spoken American English For comments on an early draft of this paper, I would like to thank H S Gopal, Sandy Thompson, and Keith Slater

English from a wide variety of speakers and speech events See Lenk (to appear) for a comparison and discussion of the transcription systems used in the SBCSAE and a number of other spoken English corpora The Du Bois et al transcription system has also been used in research on a number of other languages besides English

In general, the Du Bois et al system provides a fairly broad representation of spoken language It is not a full model or representation of prosody such as the ToBI system (cf Silverman et al 1992), and is not intended as such Again, consult Lenk (to appear) for a discussion of the goals of transcription systems for spoken language corpora

In this transcription system, the basic unit of spoken language is referred to as the IU (Intonation Unit), which is broadly defined as follows

> "Roughly speaking, an intonation unit is a stretch of speech uttered under a single coherent intonation contour It tends to be marked by cues such as a pause and a shift upward in overall pitch level at its beginning, and a lengthening of its final syllable "
> (Du Bois et al 1993 47)

Each IU of a speech event appears as a single line in the transcription For further definition, discussion, and justification of the Intonation Unit see Chafe (1980, 1994, inter alia), Du Bois et al (1992, 1993), and Schuetze-Coburn, Shapley, and Weber, (1991)

Du Bois et al (1993 52-57) have proposed two approaches to the coding of intonation contour One, Transitional Continuity, refers to the function of specific pitch patterns, and is defined as follows

> "When a speaker arrives at the end of an intonation unit, poised to continue on to the next -- or not continue -- the intonation contour usually gives a fairly clear indication of whether the discourse business at hand will be continued, or has finished This is transitional continuity the marking of the degree of continuity that occurs at the transition point between one intonation unit and the next" (ib 53)

Thus the Transitional Continuity of an Intonation Unit refers to various patterns of pitch on a specific IU, which indicate specific functions -- whether the speaker will potentially continue on to another IU, or whether the speaker could potentially stop speaking at the end of the current IU Du Bois et al propose three types of Transitional Continuity Final, Continuing, and Appeal Each (non-truncated) Intonation Unit in a spoken text is classified into one of these three types Final indicates that the speaker has reached a point of potential completion, and is marked in the transcript with a period at the end of the IU This symbol "indicates a class of intonation contours whose Transitional Continuity is regularly understood as final in a given language" (ib 54) Continuing Transitional Continuity indicates that the speaker will likely continue speaking, and is indicated by a comma at the end of the IU The third class of Transitional Continuity is Appeal, and is marked with a question mark in the transcription system and indicates a type of intonation used "when a speaker, in producing an utterance, seeks a validating response from a listener" (ib 55) In sum, these three classes are functional categories based on how prosodic patterns are generally interpreted in a given language

The second type of intonational information captured by the transcription system is Terminal Pitch Direction Unlike Transitional Continuity, which refers to a functional class of prosodic types, Terminal Pitch Direction refers to actual acoustics, as perceived by the transcriber

> "The symbols iconically represent the movement of pitch, at a critical location in the intonation unit at the end of the unit, i e the transition point from one intonation unit to the next In contrast to the symbols in the last section, which represent a certain aspect of intonational function, these symbols are designed to represent the auditory shape of the pitch movement " (ib 55)

The transcription recognizes three categories of Terminal Pitch Direction Rise (marked with a slash at the end of an IU), Fall (marked with a backslash), and Level (indicated by an underscore)

The coding of these two prosodic categories -- Transitional Continuity and terminal pitch direction – is based solely on the auditory perceptions of the transcriber For both categories, there exists no external, "correct", normative answer against which the codings of raters may be compared Even the category of Terminal Pitch Direction, which is acoustically based, is essentially defined by the analyst's perceptions of the data As Du Bois et al claim, only the acoustic features which are salient to human speakers are those which should be included in a transcription system Because they are defined as perceptual categories, the nearest approximation to assessing the "reality" of Transitional Continuity and Terminal Pitch Direction is the metric of interrater agreement, as there exists no external norm to judge perception

But exactly how adequate are the intuitive, perception-based judgements of transcribers? As of yet, there has been no published work addressing interrater reliability for these intonational categories of this transcription system Yet some linguists, including the present author, have proposed theoretical claims based on these categories, especially Transitional Continuity One well-known example is Ford (1993) who claims that English adverbial clauses which follow a final Transitional Continuity differ in their interactional function and significance from those which follow continuing Transitional Continuity If final and continuing Transitional Continuity are themselves not clearly agreed upon by transcribers, then the underpinnings of such theoretical claims become problematic Thus, I believe it important to test the "reality" of these categories, which is the aim of this present paper

This paper seeks two address two questions First, how strong is interrater agreement for each of these two categories? Second, does training have an effect on interrater reliability?

3 Methodology

In order to investigate these questions, I conducted an experiment in which I asked twelve native speakers of American English to independently code Transitional Continuity and Terminal Pitch Direction for 63 pre-assigned IUs of running discourse These ratings were then quantified to determine consistency of agreement among raters, and the codings of trained versus untrained raters were compared

The twelve raters for this study comprised two groups of six raters each One group of six consisted of people who had been extensively trained in the transcription system of Du Bois et al, by taking a 10-week transcription course taught by Prof Du Bois at the University of California Santa Barbara The second group of six raters consisted of "naive native speakers" who had not studied linguistics and had never received any formal instruction in transcription or analysis of spoken discourse These raters were simply given the brief definitions of Transitional Continuity and Terminal Pitch Direction presented above The full text for this study is an excerpt from "Actual Blacksmithing", one of the speech events in the Corpus of Spoken American English The snippet which formed the stimulus for the experiment consisted of 63 consecutive IUs of running discourse from the same speaker This text had already been segmented into IUs by Corpus transcribers, and marked for Transitional Continuity (Transitional Continuity markings were deleted before presenting the text to the raters) A sample (11 of the 63 IUs) appears below in (1) I then digitized the audio of these 63 IUs onto a computer, and segmented it into IUs according to the printed transcript I used these individual IUs to make an audio tape, which served as the stimulus for the experiment Each rater coded the data twice one 30-minute session for Transitional Continuity and another 30-minute session for Terminal Pitch Direction (the presentation order of these two sessions varied by every other rater) Raters did not have access to the codings from the first session when coding the second, and raters did not have access to the codings of other raters

(1) Sample Text

IU#	TEXT
18	the two corners
19	they go out
20	too
21	you know
22	okay your shoe's like this
23	you stretch this out
24	well then it's gonna make these go way out
25	too
26	and they get like this
27	then you have to round these back
28	you know

4 0 Quantification and Analysis

After all six raters from each group had independently coded the data, there were two databases (one for Transitional Continuity and the other for Terminal Pitch Direction) Each database consisted of 63 records (one for each IU) and twelve fields (one for each rater) A sample of this database appears below in Table 1, demonstrating raters' responses for IUs 18-28 Raters' initials are column titles, and IU numbers label the rows The first line for each IU contains the Transitional Continuity (for 'final', , for 'continuing', and ʔ for 'appeal') as indicated by each rater for that particular IU, while the second line contains codings of the Terminal Pitch Direction (/ for 'rise', _ for 'level', and \ for 'fall')

(Table 1) **Sample from Raw Database**

	TRAINED						UNTRAINED					
IU#	KS	JR	NK	AK	LO	JS	JJ	JM	CS	VJ	RM	SW
18	,	,					,	,	,	,		,
	\	\	\	\	\	_	/	\	_	\	\	\
19				,	,	,		,		,		,
	\	\	\	_	\	_	_	\	_	\	\	\
20												
	\	\	\	\	/	_	/	\	\	\	\	\
21	,	,	,	,	,	,	ʔ		ʔ		ʔ	,
	/	\	_	/	_	/	_	_	/	\	_	_
22	,	,	,	ʔ	ʔ	ʔ	,	,	,	,	ʔ	ʔ
	/	/	/	_	/	/	/	/	/	/	/	/
23	,			,	,	,	,	,	,	,		,
	\	\	\	\	\	_	_	\	_	\	\	_

24	 \	 \	 \	, _	 _	, _	, \	 \	, _	 \	 _	, \
25	, \	 _	 _	, /	, _	, _	 _	 /	 _	 _	 \	, _
26	, /	, /	? /	, \	? /	, /	, /	 /	, /	, /	? /	? /
27	, /	, _	, _	, _	, \	, _	, /	, \	, /	, \	, \	, _
28	, /	? _	, /	, \	, \	? /	? \	 _	? \	, \	? _	, \

Arraying the data in this fashion allows one to see exactly how each particular IU was coded by each rater It also readily yields information on complete rater agreement – those IUs which all raters coded identically For example, the sample data in Table 1 shows complete agreement on IU 20 (which all raters marked as Final) and IU 27 (which all raters marked as Continuing) As summarized below in Table 2, in the overall database, there are 23 instances of complete agreement for Transitional Continuity 18 Continuing, 5 Final, and 0 Appeal For Terminal Pitch Direction, there are only five instances of complete agreement (3 Level, 2 Rise, and 0 Fall) The percentages of total agreement seem to indicate that raters were far more consistent among themselves at marking Transitional Continuity than they were for marking Terminal Pitch Direction

(Table 2) **Summary of Complete Agreement Among all Raters**

Transitional Continuity	Terminal Pitch Direction
Total IUs 23	Total IUs 5
Continuing 18	Level 3
Final 5	Rise 2
Appeal 0	Fall 0
%-of-whole 36 51%	%-of-whole 7 94%

This approach is of particular interest since it targets just those IUs about which all raters agree (regardless of training) Presumably, these represent the clear, canonical cases in the database of the phenomena being investigated A direction for potential future research would entail a correlational study of these particular cases What factors are similar among these IUs which led all raters to give the same ratings? Such a study would potentially give greater insight into the cues which can be used to indicate Transitional Continuity and/or Terminal Pitch Direction By understanding these clear cases, researchers could gain insight into the less-clear cases, and could more rigorously operationalize the definitions of these two intonational categories Although interesting, such an all-or-nothing binary approach to the data does not readily lend itself to assessing overall trends of interrater agreement It is these agreement trends which can best answer the questions posed earlier is agreement among the six raters in each group statistically significant? Is there a significant difference across the two groups in terms of overall coding? In order to answer these two questions, one must evaluate the agreement for each pair of raters

4 1 Pairwise Agreement Percentages

One standard approach to the quantification of interrater agreement is known as agreement percentages These are relatively easy to compute, yet their reliability as an assessment metric is extremely inadequate, for reasons which I shall discuss below (cf Bakeman and Gottman 1986 75-77) Although inadequate as statistical indices, they are nonetheless useful for suggesting agreement trends among multiple raters, such as the case of the present study

To calculate the percent of agreement among two raters, we count the number of times the two raters agree with each other on each of the coding categories, and divide by the total number of subjects being coded (in this case subjects = 63 IUs) To illustrate, consider the Transitional Continuity coding by two of the trained raters, KS and JR, as depicted in the agreement matrix in Table 3

(Table 3) **Agreement Matrix for Transitional Continuity coding by KS and JR**

		JR			
			,	ʔ	TOTAL
KS		10	0	0	10
	,	6	41	4	51
	ʔ	0	2	0	2
	TOTAL	16	43	4	63

Agreement Percentage = 80 95%

This agreement matrix illustrates the exact numbers out of the 63 IUs which KS and JR both coded identically (the diagonal of the table), and also illustrates the number of IUs on which they disagreed, and characterizes the nature of the disagreement For example, reading down the first column of the matrix gives the following information both KS and JR coded the same 10 IUs as Final, there are 6 IUs which JR coded Final but which KS coded Continuing, there are 0 IUs coded Final by JR and Appeal by KS, the bottom cell of this column indicates the total number of IUs JR marked as Final -- 16

Looking at the diagonal of the matrix in Table 3, KS and JR both coded the same 10 IUs as Final, 41 IUs as Continuing, and 0 IUs as Appeal These are the numbers of IUs about which both raters agreed Adding these numbers, we see that KS and JR independently agreed on 51 of the 63 IUs in the database, thus showing 80 95% agreement

Since there are six trained and six untrained raters, there are fifteen rater pairs in each group, as shown by the formula $(n^2-n)/2$, where n = 6 raters Table 4 (see next page) lists the rank-ordered pairwise agreement percentages for Transitional Continuity, and Table 5 (see next page) lists the same information for the raters' coding of Terminal Pitch Direction In terms of the overall agreement percentages, we observe that raters in both trained and untrained groups were dramatically more consistent at coding Transitional Continuity than they were at coding Terminal Pitch Direction The trained raters averaged 76 40% agreement for Transitional Continuity, versus only 56 30% agreement for Terminal Pitch Direction Untrained raters averaged 73 02% agreement for Transitional Continuity and only 57 88% for Terminal Pitch Direction In terms of differences in rate of agreement between the groups, trained raters showed slightly more agreement on Transitional Continuity than did untrained raters, while untrained raters showed a seemingly insignificant higher rate of agreement on Terminal Pitch Direction than did trained raters

(Table 4) **Rank-Ordered Agreement Percentages for Transitional Continuity**

		TRAINED			UNTRAINED
	%	RATERS		%	RATERS
MAX	90 48	AKxJS	MAX	85 71	JJxCS
	80 95	KSxLO		79 36	VJxSW
	80 95	KSxJR		79 36	JMxVJ
	79 36	KSxJS		77 78	JJxVJ
	79 36	KSxAK		77 78	JMxSW
	76 19	KSxNK		74 60	JJxSW
	74 60	JRxAK		73 02	VJxRM
	74 60	LOxJS		71 43	CSxVJ
	74 60	NKxJS		71 43	CSxSW
	74 60	AKxLO		69 84	JJxJM
	73 02	JRxLO		68 25	RMxSW
	73 02	NKxLO		68 25	CSxRM
	73 02	JRxJS		66 67	JJxRM
	73 02	JRxNK		66 67	JMxCS
MIN	68 25	NKxAK	MIN	65 08	JMxRM
AVG	76 40%		AVG	73 02%	

(Table 5) **Rank-Ordered Agreement Percentages for Terminal Pitch Direction**

		TRAINED			UNTRAINED
	%	RATERS		%	RATERS
MAX	74 60	JRxNK	MAX	69 84	JMxVJ
	71 43	NKxLO		68 25	JMxSW
	66 67	JRxLO		66 67	VJxSW
	61 90	JRxJS		66 67	JMxRM
	60 32	NKxJS		58 73	JJxJM
	60 32	NKxAK		57 14	VJxRM
	58 73	LOxJS		57 14	CSxRM
	57 14	JRxAK		57 14	JJxCS
	57 14	KSxJR		55 56	JJxSW
	50 79	KSxNK		55 56	CSxSW
	49 21	AKxLO		55 56	RMxSW
	47 62	AKxJS		52 38	JMxCS
	44 44	KSxJS		52 38	JJxRM
	42 86	KSxAK		47 62	CSxVJ
MIN	41 27	KSxLO	MIN	47 62	JJxVJ
AVG	56 30%		AVG	57 88%	

As mentioned earlier, agreement percentages are a very rough measurement and do not indicate statistical significance As Cohen (1960 38) aptly states about this type of agreement index "It takes relatively little in the way of sophistication to appreciate the inadequacy of this solution " It can be useful, however, as a descriptive tool for an overall view of the data For example, this measurement was used in assessing inter-transcriber agreement for the TOBI system (Pitrelli, Beckman, and Hirschberg 1994) These researchers did not pursue statistical significance, but used an average of pairwise agreement percentages (Although their data was arrayed in a slightly different manner from the current study, the metrics used are equivalent)

The main problem with this approach is that at least some of the agreement between two raters is expected simply by chance, which agreement percentages do not take into account

4 2 Pairwise Kappas

The kappa statistic, proposed by Cohen (1960) corrects for chance Kappa provides the overall rate of agreement between two raters, once chance agreement has been removed (For a more detailed discussion of Kappa, see Cohen 1960, Fleiss 1981, Bakeman and Gottman 1986)

The first step in calculating kappa is to array the data in an agreement matrix, but using decimal proportions in the matrix cells rather than raw scores To convert the raw scores of Table 3 for KS and JR into decimal proportions, divide each cell by 63 (the total number of ratings) which yields the agreement matrix in Table 6

(Table 6)
Agreement Matrix for Transitional Continuity coding by KS and JR Decimal Proportions

		JR			
			,	ɔ	TOTAL
KS		1587	0	0	1587
	,	0952	6508	0635	8095
	ɔ	0	0317	0	0317
	TOTAL	2540	6825	0635	1

p_o = 8095
p_c = 5948
κ = 5298

This agreement matrix shows that KS and JR both coded the same 15 87% of the IUs as Final, the same 65 08% of the IUs as Continuing, and 0 IUs as Appeal To calculate the proportion of observed agreement, p_o, simply add the diagonal to get 8095 This number means that KS and JR agreed on 80 95% of the coding, and is the same as the percent of agreement discussed in the previous section

The proportion of chance agreement, p_c, is equal to the sum of the diagonal probabilities In other words, for each of the categories in this matrix, what is the probability that the codings by KS and JR were simply due to chanceɔ To calculate the probability, take the product of like marginals For example, multiply the total proportion of Finals coded by KS (1587) and the total number of Finals coded by JR (2540) to get (1587* 2540) = 0403 probability of chance agreement on final Transitional Continuity Do the same for Continuing (8095* 6825 = 5525), and for Appeal

(0317* 0635 = 0020), and add these three products together to get p_c = 5948 Now that we know the values of p_o and p_c, calculating kappa is simply a matter of plugging the numbers into the formula for kappa given in (2) below

(2) **Formula for Kappa**

$\kappa = (p_o - p_c)/(1 - p_c)$

For KS and JR we calculate the pairwise kappa as κ = 5298 In other words, the proportion of agreements between KS and JR after chance has been excluded is 5298 The process of calculating kappa is identical for the remaining 14 pairs of raters in this group, the 15 pairs of raters in the untrained group, and for both groups' codings of Terminal Pitch Direction I present the rank-ordered pairwise kappas and their averages in tables 7 (below) and 8 (see next page) From these tables we observe the same general trends we saw earlier Based on the average pairwise kappa scores for both groups, agreement is substantially higher for coding Transitional Continuity than it is for Terminal Pitch Direction There is a very slight difference in the kappa scores of the trained and untrained groups the trained raters show slightly more agreement on Transitional Continuity, while the untrained raters show slightly more agreement on Terminal Pitch Direction The averages of pairwise kappas for each group are listed in Table 9 (see next page) for ease of comparison

(Table 7) **Rank-Ordered Pairwise Kappas for Transitional Continuity**

		TRAINED			UNTRAINED
	K	RATERS		K	RATERS
MAX	7872	AKxJS	MAX	7393	JJxCS
	5568	KSxLO		5577	JJxVJ
	5298	KSxJR		5480	JMxVJ
	5168	NKxJS		5214	VJxRM
	5034	KSxNK		5108	VJxSW
	4926	NKxLO		4650	CSxRM
	4852	KSxJS		4569	CSxVJ
	4780	LOxJS		4566	JMxSW
	4773	JRxNK		4437	JJxSW
	4641	AKxLO		4273	JJxRM
	4572	KSxAK		4158	CSxSW
	4439	JRxLO		4150	RMxSW
	4409	JRxAK		3871	JJxJM
	4233	JRxJS		3697	JMxRM
MIN	3878	NKxAK	MIN	3543	JMxCS
AVG	4963		AVG	4712	

(Table 8) **Rank-Ordered Pairwise Kappas for Terminal Pitch Direction**

		TRAINED			UNTRAINED
	K	RATERS		K	RATERS
MAX	5957	JRxNK	MAX	5267	JMxVJ
	5362	NKxLO		4952	JMxSW
	4723	JRxLO		4528	VJxSW
	3923	JRxJS		4460	JMxRM
	3923	KSxJR		3666	JJxJM
	3621	NKxAK		3462	JJxSW
	3380	NKxJS		3398	CSxSW
	3285	JRxAK		3300	VJxRM
	3190	KSxNK		3255	JJxCS
	3189	LOxJS		3239	CSxRM
	2430	KSxJS		2775	JMxCS
	1966	KSxAK		2699	RMxSW
	1890	AKxLO		2434	JJxVJ
	1812	KSxLO		2291	CSxVJ
MIN	1406	AKxJS	MIN	2079	JJxRM
AVG	3337		AVG	3448	

(Table 9) **Averages of Intragroup Pairwise Kappas**

	Transitional Continuity	Terminal Pitch Direction
Trained	$\kappa =$ 4963	$\kappa =$ 3337
Untrained	$\kappa =$ 4712	$\kappa =$ 3448

When the number of raters > 2 (as in the present study), it has been demonstrated that the average of pairwise kappas is an equivalent statistical measurement to more complex metrics (Conger 1980) Thus, the averages of pairwise kappas listed in Table 9 are sufficient measures of agreement within each of the groups I will refer to these as measures of the intragroup strength-of-agreement, characterizing the overall consistency of responses of the six raters in each group

But what exactly do these numbers indicate? The upper limit of kappa is 1 0, representing complete agreement A kappa of 0 indicates agreement due to chance A negative value of kappa indicates agreement less-than chance The lower limit of kappa varies, depending on the marginal probabilities in the agreement matrix, but its value is always between -1 and 0 All the kappas in my data indicate interrater agreement greater than chance, in ALL cases

In a more rigorous statistical study, it would be necessary to compute the standard error of the intragroup strengths-of-agreement in order to arrive at a specific level of certainty about the results However, because of confounding factors (e g the lack of independence among the kappas used to compute the overall average), calculating a standard error in this case would be neither straightforward nor necessarily reliable Furthermore, since there are only 6 raters in each group, the sample size is not large enough to use general parametric measurements For these reasons, studies in the social sciences which use ratings from multiple raters tend to employ kappa merely as a descriptive statistic

Fleiss (1981) and Landis and Koch (1977 165) suggest a more general, descriptive approach to assessing the significance of kappa They give a set of benchmarks for ranges of kappa values, summarized in (3) below

> (3) **Benchmarks for Kappa as a descriptive statistic**
> "For most purposes, values greater than 75 or so may be taken to represent excellent agreement beyond chance, values below 40 or so may be taken to represent poor agreement beyond chance, and values between 40 and 75 may be taken to represent fair to good agreement beyond chance " (Fleiss 1981 218)

Applying these general assessments to the overall averages presented in Table 9, we observe that the intragroup strength-of-agreement in both groups is "fair" for Transitional Continuity and "poor" for Terminal Pitch Direction Both are a far cry from the score of 75 generally needed for "excellent" interrater agreement

The second question I raised earlier concerns the effect of training Is it possible to calculate an intergroup differential -- the difference between trained and untrained raters coding each of the two categories? As shown in table 9, the overall difference between the average of pairwise kappas is very slight the kappa for trained raters averaged 0251 higher than untrained for Transitional Continuity, and 0111 lower than untrained for Terminal Pitch Direction These differences suggest the effect of training, if present at all, is negligible There is yet another factor which must be taken into account here how do we know that trained and untrained raters are coding based on the same criteria? What if, for example, untrained raters regularly agree among themselves that IUs x, y, and z should be coded with Final Transitional Continuity, while the trained raters also agree among themselves but all code those same IUs as Appeal? In such a case, although intragroup agreement scores are the same, intergroup scores would be radically different To test for such disagreement between groups, I calculated kappa for each of the 36 unique pairs of trained and untrained raters If the average intergroup kappas for Transitional Continuity and Terminal Pitch Direction are roughly equivalent to the intragroup kappas for these same categories, then we can conclude that raters are in fact using the same criteria in coding the data Table 10 compares the average intergroup kappas with the intragroup averages presented earlier

(Table 10) **Summary of Intra- and Intergroup Average Kappas**

	Transitional Continuity	Terminal Pitch Direction
Trained	$\kappa =$ 4963	$\kappa =$ 3337
Untrained	$\kappa =$ 4712	$\kappa =$ 3448
Intergroup	$\kappa =$ 4815	$\kappa =$ 3503

We observe that the scores are very similar, and can conclude that agreement between the two groups of raters is roughly consistent with agreement within each of the two groups, giving further evidence that the effect of training, if it exists at all, is negligible

5 Summary and Conclusions

The aim of this study was to assess the "reality" of the prosodic categories of Transitional

Continuity and Terminal Pitch Direction as proposed by Du Bois et al. for the transcription of spoken language. The results demonstrate that interrater agreement for both Transitional Continuity and Terminal Pitch Direction is greater than chance, lending support to the existence of these categories. However, as demonstrated by the overall average of pairwise kappas, interrater agreement is rather low: "fair" for Transitional Continuity (κ = .4963 for trained raters and .4712 for untrained) and "poor" for Terminal Pitch Direction (κ = .3337 for trained raters and .3448 for untrained). The data also show that both trained and untrained raters are substantially more reliable at coding Transitional Continuity than they are at coding Terminal Pitch Direction. The data also suggest that the effect of training is negligible.

Do these results indicate that the categories of Transitional Continuity and Terminal Pitch direction should be abandoned? I believe not. As shown by the average pairwise kappas in the previous section, coding of these two categories is greater-than chance, suggesting their validity for the raters. Rather than abandoning them outright, I believe the study indicates a need for more rigorous operationalization of their definitions. The remainder of this paper will focus on potential future research with this goal in mind.

For both categories, one means of achieving a better understanding of the cues transcribers use in their identification is to examine the clear cases where all raters agreed. As discussed in section 4.0 and as summarized in Table 2, roughly 36 percent of all IUs received unanimous agreement by all raters for Transitional Continuity, and roughly 8 percent of all IUs for Terminal Pitch Direction. These should be taken as clear, unambiguous instances of the two categories. Future research is warranted to examine the cues which raters used in their coding of these particular IUs. By understanding what the raters were doing in these clear cases, we can achieve a more stringent operationalization of the definitions of these categories which can then be applied to the less-clear cases.

For Transitional Continuity, I believe future research should focus on the identification of other Transitional Continuity types, beyond the three defined by Du Bois et al. Perhaps some of the disagreement in coding arises from the conflation of several distinct Transitional Continuity contours into only three categories. Perhaps given more options, raters might be more consistent at coding the "in-between" cases. For example, Genetti and Slater (to appear) have identified six distinct Transitional Continuity types in elicited narratives in Dolakhae Newar (a Tibeto-Burman language spoken in Nepal): prototypical final, narrative final, interrogative final, exclamatory final, anticipatory continuing, and non-anticipatory continuing. While Genetti and Slater do not claim these 6 categories are valid cross-linguistically, I believe a similar approach could be useful for English as well.

There are several issues surrounding the coding of Terminal Pitch direction which very plausibly lead to the extremely low rate of agreement seen in this study. I should stress that this category is not generally used by discourse researchers, and it was not coded for in the Corpus of Spoken American English. First, as defined by Du Bois et al., Terminal Pitch Direction refers specifically to coding the direction of pitch movement at the ends of IUs. Unfortunately, Du Bois et al. do not define what is meant by the "end" of an IU--final syllable, final vowel, final word, etc. Several of the raters specifically commented on this after the experiment. Secondly, psychoacoustic research needs to be brought to bear on this issue: how reliable are subjects at perceiving pitch changes in spoken language, and what kind of training can improve this? One worthwhile direction for future research along these lines would be a comparison of transcribers' coding of Terminal Pitch Direction with the actual physical measurement of pitch slope. Thirdly, I suggest that interrater reliability would improve dramatically if transcribers were provided with pitch traces of the IUs being coded (cf. evaluations of interrater agreement for the ToBI system discussed by Pitrelli, Beckman, and

Hirschberg 1994) And, finally, as suggested for Transitional Continuity, perhaps the transcription system needs to be expanded to allow for more than only three types of Terminal Pitch Direction contours

In conclusion, while there are reasonable statistical grounds to assume the existence of the two categories discussed in this paper, relatively low interrater agreement suggest the need for further operationalization of their definitions It is hoped that the methodology outlined in this paper will be employed in the justification of other linguistic coding categories as well Such empirical validation will serve to clarify many of the categories we use in language research, and will enhance the effectiveness and accuracy of the theories built upon them

References

Bakeman, Roger and John M Gottman 1986 Observing interaction an introduction to sequential analysis New York Cambridge University Press

Chafe, Wallace L 1980 The deployment of consciousness in the production of a narrative In Wallace L Chafe ed , The pear stories cognitive, cultural, and linguistic aspects of narrative production, 9-50 Norwood, NJ Ablex

Chafe, Wallace L 1994 Discourse, consciousness, and time the flow and displacement of conscious experience in speaking and writing Chicago University of Chicago Press

Cohen, Jacob 1960 A coefficient of agreement for nominal scales Educational and Psychological Measurement 20 37-46

Conger, Anthony J 1980 Integration and generalization of Kappas for multiple raters Psychological Bulletin 88 322-328

Du Bois, John, Stephan Schuetze-Coburn, Danae Paolino, and Susanna Cumming 1992 Discourse transcription Santa Barbara Papers in Linguistics Vol 4, University of California at Santa Barbara

Du Bois, John, Stephan Schuetze-Coburn, Danae Paolino, and Susanna Cumming 1993 Outline of discourse transcription in Jane A Edwards and Martin D Lampert, eds , Talking data transcription and coding methods for language research, 45-89 Hillsdale, NJ Lawrence Erlbaum

Fleiss, Joseph L 1981 Statistical methods for rates and proportions, 2nd ed New York Wiley

Ford, Cecilia E 1993 Grammar in interaction adverbial clauses in American English conversations Cambridge Cambridge University Press

Genetti, Carol and Keith Slater to appear An analysis of syntax/prosody interactions in a Dolakhae Newar rendition of the Mahabarata In George van Driem, Ed Himalayan Linguistics Berlin Mouton de Gruyter

Landis, J Richard and Gary G Koch 1977 The measurement of observer agreement for categorical data Biometrics 33 159-174

Lenk, Uta to appear Notation systems in spoken language corpora Handbook of Pragmatics

Pitrelli, John, Mary Beckman, and Julia Hirschberg 1994 Evaluation of prosodic transcription labeling reliability in the ToBI framework in Proceedings of the International Conference on Spoken Language Processing 3 123-126

Schuetze-Coburn, Stephan, Marian Shapley, and Elizabeth Weber 1991 Units of intonation in discourse a comparison of acoustic and auditory analyses Language and Speech 34 207-234

Silverman, Kim, Mary Beckman, John Pitrelli, Mari Ostendorf, Colin Wightman, Patti Price, Janet Pierrehumbert, and Julia Hirschberg ToBI A standard for labeling English prosody in Proceedings of the International Conference on Spoken Language Processing 2 867-870