TRANSITIONS TO TERMINALS, UTTERANCE TO SENTENCE,?...: DISCREPANCIES AMONG JUDGES ASSIGNING ORTHOGRAPHIC SENTENCE BOUNDARY MARKERS TO TRANSCRIPTIONS OF NATURAL SPEECH EVENTS

Harriett Nutt Hays
Columbia, Missouri

It would seem to be a simple matter for anyone with normal hearing to listen to a conversation, a monologue or any but noise-ridden utterances, and to set down what was heard, verbatim, on a piece of paper.

Evidence from pilot studies conducted over the past two years, however, indicate that a comparison of two or more interpretations by participants or observers of the same linguistic event will probably yield numerous discrepancies (H. Hays 1968, 1969, 1970). The most pertinent of these for linguists can be classified into general categories of form, grammar, ostensible lexical reference, intended and received 'actual' meaning. Elements of the language system are so multi-functional that one discrepancy might be analyzed as a combination of categories, another discrepancy as just one of them; and any discrepancy might be considered trivial or significant according to the focus of the analyst.

Most differences in interpretation (potential errors or distortions) are signalled in the linguistic event by some difference in form whether that form be manifested in the tonal, amplitudinal, durational, qualitative, grammatical, alphabetic, numeric, or other pertinent dimension of linguistic symbolism. For some analysts most differences in form will not be crucial unless a definite portion of the semantic spectrum is affected. For others, even a slight formal variation could be significant, affecting interpretations of grammatical structure, style, general statistical measure, or nuance of meaning.

This paper examines just the <u>non-special</u> punctuation discrepancies among two sets of multiple transcriptions of the same event prepared by a number of processors, some of whom had training in linguistic observation and some of whom did not. (Non-special symbols are, for this analysis: a series of dashes (----) used to indicate an indistingishable utterance; comma (,); three dots (...); three dots plus hash mark (...#); question mark (?); period (.); exclamation point (!); and capital letter. Special symbols include all other punctuation marks, and are not discussed here.)

The manuscripts prepared were intended to be verbatim representations of two seven-minute segments of videotaped recordings, of fairly decent quality, of two different sixth grade public school classrooms, and were the first sets in a series of experiments on producing accurate transcriptions of ongoing verbal events into standard orthography (H. Hays 1968, 1969, 1970). It was important for these particular transcriptions to be faithful to the verbal stream since a large body of

replicable empirical research on interpersonal interaction, gross activity patterns, features of linguistic style, and actual meaning of utterances was to be performed on many classroom tapes recorded for that purpose. Since the boundaries of the analysis units for the different investigations were not expected to be co-terminous, it was planned to link or cross-reference the separate analyses by means of annotations of real time on the tape and by enumeration of the sentences of the manuscripts.

It is known that standard orthographic rules for punctuation contain options for the user. Therefore, it was predicted that there would be some variation among versions for internal punctuation, especially in use of commas, and that there might be some discrepancy in assignation of periods and question marks. It was not predicted by most of the potential analysts of the classroom materials that there would be extensive discrepancy among transcriptions for either word or sentence boundary interpretation (H. Hays 1968a,b, 1969b).

As it turned out, there were many discrepancies among manuscripts, discrepancies which ranged from apparently trivial typographical errors to complete (and probably significant) omissions or disagreements for individual words and entire phrases. Eight to thirteen-way discrepancies were found for some word and phrase interpretations, but no more than five-way differences were attested for a single punctuation mark position.

The discrepancies were examined for processor agreement and situation of context, but it was determined that neither majority opinion nor contextual match could be considered reliable indices for predicting the probable 'correct' version for these segments. Although both the trained linguists and the trained social scientists who worked on the manuscripts tended to make more detailed transcriptions (particularly regarding annotations of the behavior) than did persons untrained in observing communicative behavior, definite and potentially significant discrepancies were also found among their versions of word and punctuation units.

The nature of the discrepancies found among the sets of experimental materials illustrating today's paper, a discussion of the recording and processing conditions, of the event, of the personnel and of the analysis procedures used was reported in H. Hays 1970a.

Experimental design

On the same stereophonic playback equipment having a mechanism for forwarding and reversing the tape an infinite number of times, at the user's discretion to perceive acccurately what transpired, five original transcriptions (A,B,C,D,F) were made separately for each of two seven-minute segments (I and II) of videotaped recordings. These transcriptions were compared for accuracy against the recording by a number of post-editors, who made additions, deletions or modifications of the texts according to their own perceptions, taking care that all public utterances were reproduced without improvements upon their grammaticality or beauty. Standard orthography was used, and all vocal interjections, false starts, incomplete utterances (often ignored in transcriptions)

were to be included. Final editors compared all previous graphic versions against the videotape and inserted, deleted or modified those transcriptions which seemed inappropriate.

From the raw transcripts for Set I a handwritten transcription (E) was calculated, based upon majority agreement of processors and contextual fit. This manuscript also was edited. The E transcription for Set II was not handwritten in manuscript format, but remained in chart form, aligned with all versions of II, A-D. This chart was then edited, with the 'correct' version checkmarked by the editors. All manuscripts were aligned with all calculations to match all versions for I and all versions for II. The resultant charts (G) were organized for Universal and Discrepant Processor Agreements for Emitter (sources of utterance), Annotation (disambiguation of utterance by processor) and Utterance itself (including punctuation). From the final manuscripts a linear alignment was made in preparation to coding and entering all of the contrasted information into a computer for tabulation of contingency variables. (This input has not been effected yet.) The Summary of Manuscript Characteristics (Figure 1) will give a general indication of the amount of effort expended on producing the transcriptions. Constraints on punctuation conventions are discussed below.

Measurements

Eye measurements and hand tabulations of discrepancies were made for gross manuscript characteristics and discrepancy sequencing. Discrepancy classes were established heuristically and tabulated for Punctuation (internal and external, special and non-special), Word, Phrase, Emitter, Spelling, Annotation. These were examined for Omission (lack of recognition of a unit asserted by one or more processor) and Interpretation (variation in form or referent), and Processor Agreement.

Figure 2 reproduces the first few lines of each raw (unedited) transcript for Set I and the last edited version of its last calculated manuscript. Figure 3 gives the gross manuscript characteristics for each of these samples; and Figure 4 charts all versions (edited and unedited) of the tape segment represented by the last lines of the samples in Figure 2, indicating processor source for each option. Figure 4 reflects the generally strong influence that previous graphic versions tended to have on the perception of the editors, but note in Figure 11 that this did not always hold true.

Implications

Detailed examination of each class of discrepancy among the transcriptions and consideration of discrepancy significance, has uncovered a number of interesting patterns which suggest that there is much room for further thorough and serious investigation, not only of punctuation problems, but of other variations in the conversion of oral English to standard orthography. It suggests that there may be a greater qualitative difference between the systems of speech and writing than is generally assumed, and that conversion from one to another contains translation problems which may be similar to those associated with other source-target language conversion (H. Hays 1969b, 1970a,b,c,d,e).

Manuscripts	A	B	C	D	E	F	G	A	B	C	D	E	F	G
Transcribers	a	b	c	d	e	b	s	a	b	c	d	s	b	s
Pages	2.2	2.2	2.5	2.5	11	4.5	87	4	5	7	5		4.5	118
Lines	76	72	81	84	101		727	71	61	63	76		63	770
Words	548	541	525	562	664	557	640	560	496	474	548		550	645
Sentences	84	83	87	81	81	67	81	67	71	73	74		74	183
	(Universal words: 326)							(Universal words: 305)						
	(Universal sentences: 9)							(Universal sentences: 30)						
Post-Editors	t	r	q	o	e	e	g	f	m			s	e	s
	t	r	q	o	g	f	g		m			g	f	g
	e	f	p	l	j	f	n		n				g	g
	e	f	p	l		n	n		n				k	v
				g										g
Total Processors	3	3	3	4	3	4	3	2	3	1	1	2	5	5
Total Verifications Against Tape	5	5	5	6	3	5	2	2	5	1	1	2	5	1
Minimum Processing Time (in minutes)	120	130	130	135	550	180	999!	140	695	40	40	100	395	999!

Figure 1. Summary of Manuscript Characteristics for Sets I and II

Emitter	Lines	Transcriber A
1.	1	(T) All of us, yes. //Have all of you all completed the reading on page 109?
	2	All right let's discuss how a friendly letter can be like having a visit with
	3	someone. What are some of the that we a talk about in a friendly letter?
	4	What's the atmosphere of a friendly letter?
2.	5	(?) You mean the parts.
3.	6	(T)= No not the parts, honey, the atmosphere. David.
4.	7	(David) The attitude of the person who writes.

		Transcriber B
1.	1	(T) //Have all of you all completed the reading on page 109?
2.	2	(Class) Yes
3.	3	(T) Oh right let's discuss how a friendly letter can be like having a visit
	4	with someone. What are some of the things that we a talk about in a friendly
	5	letter? What's the atmosphere of a friendly letter?
4.	6	(?) You mean the parts?
5.	7	(T) No not the parts honey the atmosphere. Doyle?
6.	8	(Doyle) You mean the attitude of the person like.

Parentheses enclose emitter designates.
Slashes enclose annotations.
// = common point of origin.
--- = indistinguishable utterance.
= analyst questions.

Raw A,B,C,D = Control
Raw F done by B transcriber
Raw E and G calculated from preceding graphs

For reasons of space and clarity, the text sample does not portray any edited version, except for G.

Figure 2. Reproduction of the First Few Lines of All Unedited Transcriptions (Original and Calculated) of Set I, and the Last Edited Version (N) of the Last Calculated Transcription (G).

Emitter	Lines	Transcriber C
1.	1	(T) _____. //Have all of you completed the reading of page 109.
2.	2	(Class) Murmurs.
3.	3	(T) Allright, let's discuss how a formal letter can be like having a visit
	4	with someone. What are some of the things that we uh talk about in a formal
	5	letter? What's the atmosphere of a formal letter?
4.	6	(?) _____
5.	7	(T) No, not the parts, honey, the atmosphere. Dale.
6.	8	(Dale) The attitude of the person writin it.

Transcriber D

1.	1	(T) All right, have ya. . . Yes? //Have all of you all completed the readings
	2	on page 109?
2.	3	(Class) Yes
3.	4	(T) All right, let's discuss how a friendly letter, letter can be like having
	5	a visit with someone. What are some of the, uh, things that we, uh, talk about
	6	in a friendly letter? What's the atmosphere of a friendly letter?
4.	7	(?) The, uh, parts.
5.	8	(T) No, not the parts, honey, the atmosphere. Dana?
6.	9	(D) The attitude of the person you write.

Transcription F (Transcriber B)

1.	1	(T) //Have all of you completed the reading on page 109?
2.	2	(Class) /in unison/ Yes.
3.	3	(T) Oh right let's discuss how a friendly letter can be like having
	4	a visit with someone. What are some of the things that we talk about
	5	in a friendly letter? What's the atmosphere? of a friendly letter?
4.	6	(---) /girl/ You mean the parts?
5.	7	(T) No not the parts honey, the atmosphere. Darla.
6.	8	(Darla) The attitude the person writes.

Figure 2 (continued)

Transcription E (Calculated from Edited A,B,C,D)

Emitter	Lines	
1.	1	(T) All right have ya... Yes? /Looks at student who has raised
	2	her hands/ (S) /Shakes her head negative///Have all of you all
	3	completed the reading on page 109?
2.	4	(Class) /Murmurs/ Yes.
3.	5	(T)= All right, let's discuss how a friendly letter, letter /enunciates/
	6	can be like having a visit with someone. What are some of the, uh,
	7	things that we, uh, talk about in a friendly letter? What's the
	8	atmosphere of a friendly letter?
4.	9	(?) You mean the parts?
5.	10	(T) No, not the parts, honey, the atmosphere. Dale?
6.	11	(Dale) You mean the attitude of the person writin it.

Transcription G (Calculated from All Edited Versions, A,B,C,D,E,F)

1.	1	(T) Have all of you all completed the reading on page 109?
2.	2	(Class) /a few students/ /sync/ /murmurs/ Yes.
3.	3	(T) Alright, let's discuss how a friendly letter, letter /enunciates/
	4	can be like having a visit, visit with someone. What are some of the
	5	uh things that we ah, talk about in a friendly letter? What's the
	6	atmosphere, of a friendly letter?
4.	7	(?) Um You mean the, uh, parts?
5.	8	(T) No, /annoyed/ not the parts, honey, the
	9	atmosphere. Daryl?
6.	10	(Daryl) You mean the uh, attitude of the person writing.

Figure 2 (continued)

Emitter	Lines	Final Version of G (last edited version against the tape)
1.	1	(T) Have all of you all completed the reading on page one oh nine?
2.	2	(Part of Class) /a few students/ /softly/ /syncopated/ /murmurs/
3.	3	(T) Alright, let's discuss how /enunciated/ a friendly letter
	4	/mispronounced ledder/, letter /enunciates/ can be like having a visit
	5	visit /high/ with someone. What are some of the uh, things that we,
	6	umm /short pause/ talk about in a friendly letter? What's the
	7	atmosphere, /slight break/ of a friendly letter? /enunciates/
4.	8	(---) /Boy/ ---- /three words sound like big and sweet/
5.	9	/Girl/ Um You mean uh parts?
6.	10	(T) No, /annoyed/ not the parts, honey, /slightly annoyed/ the
	11	atmosphere. David?
7.	12	(David) You mean uh attitude of the person like?

Figure 2 (continued)

Transcript	Lines	Sentence Punctuation: Initial	Medial	Terminal	Words	Emitter	Annotations
A.	7	8	2	8	67	4	0
B.	8	9	%	8	70	5	0/1
C.	8	8	3	8	63 + --	5	0
D.	9	9	8	8	70	5	0
E.	11	9	7	9	73	6	2
F.	8	9	1	10	65	6	2
G.	10	10	7	9	74	6	5
G(N).	12	9	6	7	74 + --	6	14

Key: -- Indistinguishable utterance (from 1 to 9 words in length)
(N) Identity of final editor of Transcript G of Set I

Figure 3. Gross Manuscript Characteristics For the First Few Lines of Text of the Five Original Transcripts (A,B,C,D,F), The Two Calculated Transcripts (E,G), and The Last Version of Transcript G Edited Against the Tape.

Darla	Fb,Fe	/Dale/	D	.	A,C,F	(Darla)	Fb,Fe
Doyle	Bb,Br	%	R	?	B,D,E,G	Doyle	Bb,Br
David	A,G					David	A,G
Dale	C,Bf,E					Dale	C,Bf,E
Dana	Dd,Do,Dg					Dana	D,Do,Dg
Daryl	Ff^1,Ff^2,Fh					(D)	Dl
D---	Dl						

You mean	E,B,G	the	B	uh	Eg,Dg,G	,	Eg,Dg
%	R	The	R	%	R	%	R
		%	G				

attitude	U	%	F	the person	U	who	A
		of	G			you	D
						%	B,C,E,F,G

writes	A,Fb	/rait.n/	Dg	it	C	?	Dg,G
%	B,C,G	%	R	like	D,G	.	R
writin	E			%	R		
writing	Fe,Dg						
write	Dd,Do,Dl						

Key:

-- Indistinguishable utterance (anywhere from one to 9 words in length)

% Lack of recognition of unit by one or more processors of unit attested by at least one processor of Set I.

U Universal agreement (attestation of form and referents) among processors.

R All the rest of the processors not explicitly listed by the upper/lower case identity code for processors in Figure 2.

Figure 4. Alignment of All Versions (Raw and Edited) and Processor Source for Final Portion of Set I, Illustrated in Figure 2

Problems

Some of the difficulties encountered in analyzing and classifying the discrepancies may be related to such translation phenomena. For punctuation, the most obvious difficulty lies in ascertaining the meaning of the symbols used. The requirements of written discourse that all utterances be segmented by terminals which enclose strings of supposedly specific structures, including what are referred to in standard school grammars as 'subjects', 'predicates' and 'complete thoughts', may confound the problem. Spoken discourse, particularly with informal style, is interspersed with what would be considered 'fragmentation' in written grammatical tradition. The problem of representing these 'incomplete' thoughts is difficult for translators who have been given no guidelines, particularly when they have differing views of 'completeness' or 'grammaticality' (Hays 1969b, 1970a,d,e).

A combination of inherent orthographic rule inconsistency (options allowed to the user in handling commas and terminals, for instance) with the multiple referentiality potential of most symbols can make translation of the oral stretch and analysis of the intent of translator and speaker a messy business. (Seemingly clearcut orthographic rules exist for marking some questions having 'diagnostic' forms of question word or inverted verb, but the diagnosis would be incorrect if it were applied to exclamatory utterances like (What a beautiful day! Is that a fact!), the cues for which appear to be drawn from the prosodic referent of the speech stretch.)

Complicating the problems of rule inconsistency is the difficulty of distinguishing actual discrepancies from agreements masked by discrepancies for other categories. An agreement on sentence initial boundary, for instance, can be obscured by a disagreement on the form which signals the boundary. See below, Figure 5. Examples 6,7,8,9, Figure 6 and Figure 7 which reflect discrepancies signalled or masked by upper-lower case disagreement. To offset this, in the analysis procedure, capitalness is now abstracted as a feature of punctuation, and separate from the quality of the symbol which carries it (H. Hays 1970a,b,c).

Akin to this type of unmasking is the handling of 'zero-utterances' (represented in notation here as (%) and indicating lack of recognition of one or more forms selected by one or more processors), and of Indistinguishable utterances (represented in notation here, and in the manuscripts, by a series of dashes (---) acknowledging the presence of an utterance, the character of which is indeterminable by the processor in question.) It is not self-evident, for instance, whether a % form is an omission, an importation, or an accurate rendition of the situation (H. Hays 1968, 1969, 1970).

There are at least three types of physical barrier, then, to the recognition of discrepancies: inherent inconsistencies in orthographic punctuation rules; masking of agreements by discrepancies of other types (both problems discussed above) and, perhaps the most frustrating difficulty of all, the lack of a one-to-one translation system for oral symbol units into graphic symbol units.

What is represented, for instance in the physical speech medium by the quantitative indicators (prosodic features) of amplitude, frequency, rate, and duration, which occur simultaneously with the qualitative indicators (consonants and vowels) are represented in writing as two-dimensional graphic signs (punctuation marks) which usually appear sequentially to the qualitative symbols (letters). The only simultaneous indicators for standard formal orthographic representations are capital letters, italics, underlines, boldface print, and the like (H. Hays 1970a). Since there is not a one-to-one ratio of the symbolic systems there, confusion is inherent in a transliteration, especially if, as is the custom, some of the signals of one (prosodic intensifying morphs, for example) are systematically excluded in the other. An acceleration or amplification of an utterance, for instance, is not represented in most transcriptions (although acceleration can be represented legitimately by omitting spaces; and amplification can be by enlarging type size.) Tonal information can be indicated in the standard orthographic rule system only by parenthetical annotation. (See annotations for teacher annoyance in Figure 2, final version of G.)

The semantic or referential units in the two systems are unlikely to be translated on a one-to-one basis if at least some of the morphological information (speaker truth value judgment from tone, for example) is omitted in transliterating from one to the other.

Scope of this paper.

This paper will examine just the physical, formal differences among processor graphic versions of the oral reality. Note that the actual oral symbolization remains undetermined for the source tapes, as it will remain for any such stretch until a more reliable form of conversion and analysis is established. (A narrow phonetic transcription and interpretation is not sufficient, for it will also be subject to similar variations among judges; as Kurath (1939, Ringaard (1964), and Lieberman (1965) have attested). Please remember, therefore, that the illustrative figures here are dependent upon orthographic parameters, with no assurance that even the last edited version for each set would be the canonic orthographic representation for the oral stretches involved. See, for example, in Figure 2, the analyst's questions regarding the editor's non-inclusion of two forms. The editor (N) was the 16th processor for this set of manuscripts.

Control Group

In spite of the small total sample, there was an attempt to retain one control group of lay transcriptions which might be representative of those from which some kinds of behavioral evidence tends to be drawn. To this end, instructions to transcribers for the raw (unedited) transcriptions of I (A-D) deliberately did not include new constraints on punctuation. The transcribers were asked to follow the procedures already established for other tapes of the same videotaped classroom series which they had already transcribed and for which there was standardized only the following general format:

All senders of utterances (Emitters) were to be designated in paren-

theses left-justified on a new line. All public utterances were to be transcribed into standard orthography. Indistinguishable utterances were to be represented by a line whose length might or might not have an impressionistic relationship to length of utterance.

Findings

Discrepancies among these versions were much heavier than expected, both by the transcribers and by many of the potential analysts (Hays 1968a,b). As expected, however, multiple purpose usage for most marks was found among the four raw transcriptions, and confusion of usage or of interpretation was attested.

See the list of examples in Figure 5 for illustration: Commas were variously inserted between phrases (Ex. 1,3,4,6,7,8), being used apparently for disambiguation as well as for pause representation. Periods and question marks were used sometimes interchangeably (Ex. 7,9,4). Upper case letters for sentence initiation and proper name signalling were sometimes confused (Ex. 9). Three dots (represented in the figures as two dots (..) for conservation of space) indicated either pause or incompletion, but it was difficult to tell which (Ex.7).

Later Instructions to Processors

The raw transcriptions of Set I, E-G, later editings of I, A-D, and all of set II were processed for the major study after a two-hour session with transcribers, and written communications were distributed to transcribers and editors to establish the following additional conventions (Hays and Hays 1969; H. Hays 1968c, 1969a,b,c).

Annotations or disambiguations clarifying the context of ambiguous utterances were to be inserted in slashes (on the assumption that, if properly marked, they could easily be left out). Punctuation was to be reduced to a minimum where possible: for terminals, only period, indicating a statement neutral in tone or feeling; question mark, indicating a definite question contour; and three dots to mark a suspended or unfinished oral sentence, were to be used. Quotation marks were to enclose matter being read out loud by an Emitter. Underlines were to signal acoustic emphasis.

Still Later Editing Conventions

For Class II, F and G, the symbol three dots plus hash mark (...#) was to indicate unfinished sentence; three dots (...), unfinished word or phrase; plain hash mark (#), unfinished word or syllable. Pauses, characterized as short, long or very long, were to be noted in parentheses.

Effect of instruction revision was clear only in the nonuse of exclamation point by all but I, A-D transcribers, and in the use in II, F and G of the symbol (...#). (It was unclear whether that symbol was used according to instructions, however). A sample of two seven-minute sets is too small, obviously, to make clear determinations of individual processor intent or habit.

The fact that there were proportionately more comma positions for Set II (107) than for Set I (77), is likely to be due to a function of speaker style and lesson content rather than the result of differences in processor instruction, although, again, the size of the sample prohibits clarification of this. Class I was a lesson in composition; Class II one in geography, where list structures were common.

The only universal internal agreement, as it happens, separated polymorphic items in a list structure, and was from Set II.

Ex. 10. Many people in Malaya, the East Indies , R (rest of processors)
" " " % D,F

As Figure 8 indicates, this was one out of a total of 20 attested permutation choice sets, 184 potential comma positions (77 for Set I, 107 for Set II). Note again that the figures do not include any orthographically potential positions other than those indicated by the processors of these two sets of seven raw manuscripts, 6 transcribers, 19 editors (a total of 16 processors and at least 34 verifications against the tape for Set I; and 12 total processors and at least 23 verifications against the tape for Set II). An immediate inference is that rules governing comma placement are a good deal less explicit than expected.

Processor punctuation style preference may have been another variable (as it might be for any symbol selections). From the seven-minute statistics it does seem that Transcriber D and Editor N did tend to use more commas than the other processors did. Since pause notation was not developed fully for these sets, however, it is also conceivable that D and N were more acutely observant of and were therefore signalling nonregular breaks in the stream of speech. Commas and annotations in the sample text of I,G's final editing (Fig. 2), compared to those of other interpretations, seem to support this.

Tentative predictions for comma representation of speech boundaries are that agreement will be slight among processors operating under standard rules of punctuation. It may be that if use of commas for disambiguation were clarified, and if a systematic notation were developed for pause length and function, agreement would be much more likely.

Although some differences in terminal markers were predicted to exist, it was not expected that there would be great variation in the placement of the markers or in interpretation of sentence or utterance boundaries (Hays 1968a,b,1969b). Note from Figure 8 that of the 28 attested terminal inclusive set choices (out of 35 total permutation sets), 18 sets (91 positions: 31 for I, 60 for II) question whether the boundary should be internal or external; 12 sets (161 positions: 85 for I, 76 for II) have clearcut agreements on placement of terminals, and only 5 sets (105 positions: 40 for I, 65 for II) agree on placement of internal markers. Of these, there were two sets (65 positions: 34 for I, 31 for II) of universal agreements for terminals, and, as mentioned above, only one instance of one set for internals (Example 11).

The character of sentence initiation could be significant for at

	Type	Examples Below
Punctuation Usage		
	Upper Case/Lower Case	
a	Proper Name	2,9
b	Heading	1
c	Sentence Initiation	1,2,5,6,7,8,9
	Markers	
d	Internal/External	
d	,/?	1
e	,/.	6
	Internal/%	3,4,8
f	%/,	
g	%/,/...	7
	External/%	
h	%/?	5
i	%/...	
j	%/./?	4,7,9
	External/External	
k	./?	7

	Type	Examples Below
Spelling		
l	Proper Name	1,3
m	Interjection	6
Interpretation		
n	Word/%	1,3,4,5,6,7,8,9
o	Word/Word	4,5,6,9
p	Phrase/%	2,8
q	Phrase/Phrase	2,3,8

Figure 5. Some Discrepancy Types Found Among Unedited Transcriptions I, A-D

	A	B	C	D	Above Types
1.	letter to Teresa's grand-father, Shakinnea?	letter? to Teresa's Grand-father? Chickenya?	letter to Teresa's grandfather Shakineen?	letter to Teresa's grandfather?	b,c,d, l,n
2.	(Class) She didn't capitalize East.	(Class) She didn't start it on the East.	(Class) Capital letter on east.	(Class) She didn't capitalize "East".	a,c,p q
3.	Now we know why Billy is writing this up so high don't we?	Now we know why Bailey's writing is set up so high don't we?	Now we know why Billy's writing this up so high don't we?	Now, we know why Billy's writing this up so high, don't we?	f,l,n, o,q
4.	her address the top.	her a-a- address? at the top?	her address at the top	her, uh, address at the top.	j,n,o
5.	called? What is the greeting?	called or what is the greeting?	called? Or what is the greeting?	called? What is the greeting?	c,h,n, o
6.	Allright. Any-thing else.	Oh right. Any-thing else.	All right, some-one else.	Alright, anything else.	c,e, m n,o
7.	oh. Ah June Butler.	of? Um-m-m-m June	Oh? Uh...June Butler.	oh? Uh, June Butler?	c,j,k,n g
8.	Some of the things that are experiences.	It tells of some experiences.	Things that tell experiences.	Uh, telling, tell-ing things of, telling your experiences.	c,f,o n,p,q
9.	David	Doyle?	Dale.	Dana?	j,b,a, n

Figure 5 (continued)

least some terminal marker selection, but for this sample there was not a clear indication that it was or was not. Sentence punctuation statistics, of course, can be highly misleading regarding agreement for initial boundary recognition. There are at least two situations in which discrepancies for other categories mask agreement in the boundary interpretation. One of these is where the discrepancy lies in the interpretation of the sentence-initiating word rather than in the acknowledgement of a new sentence. In Fig. 5, Ex. 8, each of the four sentences begins with a different word and a different symbol (Some, It, Things, Uh), yet each is capitalized and carries the requisite sentence-initiating symbol, and actually agrees with other processor boundary marks. In Ex. 5, however, only three processors agree on the boundary: B uses lower case and no separating punctuation (or), but the word and symbol quality is identical to C's initial in the upper case (Or) and different from quality of both the initial and the word (What) of A and D.

A similar masking of agreement regarding function of capital letters takes place when a proper name is designated, but the names transcribed by the processors differ from each other. In Ex. 9 (David Doyle? Dale. Dana?) there is universal agreement for capitalness, but not for quality of the word.

It is interesting to note that proper names is the other area for which, as a result of these investigations, there is predicted to be a very high frequency of observer discrepancy (Hays 1970a,c). The relatively context-free association of the names along with their grammatical or indicative function which differentiates them from most other nouns, may contribute to the discrepancy probability among the names and the punctuation surrounding them. Since many proper name designations in the current sample are discrepant for name interpreted, there are apt to be a number of cross-references or masking of agreement and discrepancy in some of the frequency figures, which are bound to represent more than a total of the actual forms harboring the discrepancies.

See Figures 6 and 7 which illustrate the frequency of type-size discrepancies for Sets I and II and the character of attested sets for Set I.

Figure 8 indicates that for manuscripts I and II of the 35 attested Choice Sets the inventory of potential punctuation symbols attested by processors for a specific position 26 have % as a choice, 6 have --, and 6 have both % and --. These potential positions were calculated from processor choices and not from the potential total positions allowed by the presumably standard orthographic rules found in the general references (Doris and Miller, and Webster's 7th) used by the office staff where the transcriptions were produced, for whom the recordings were made, and by whom the secretaries (Transcribers A,B,C,D,F) and other processors were employed (Center for Research in Social Behavior, University of Missouri, Columbia, Missouri).

How these positive non-inclusions of punctuation symbols, and positive non-identifications of utterance, punctuated or not, should be handled to differentiate them from specific symbols included in the

Set	Sentence Position: Initial Universal	Initial Discrepant	Initial Total	Internal Universal	Internal Discrepant	Internal Total
I	21	95	116	12	17	29
II	25	58	83	61	58	119
I+II	46	153	199			

Figure 6. Processor Agreement for Size of Type for Forms Attested in Sets I and II

Initial	Internal
Non-proper-names	Non-proper-names
T/%	T/t
T/T/%	T/t/t
T/%/--	T/t/%
T/X/%	Proper names — Apparent sex
T/X/--	T/T — m
T/X/--/%	T/T — f
T/T/T/X/X/X/%	T/T — m/f
T/t	T/T/T/T/T/T/T — m/f
T/t/%	T/X — m/f
T/t/t/%	T/X/Y/--/% — m
T/t/t/%/--	T/T/X/X/Y/Z/M — m/f
T/T/t/t/%	
T/t/x	T/T/T/T /T/ s/s — m/f/m/neither
T/T/t/x	
T/t/X/m	
Proper names — Apparent sex	
T/T — m/f	

Note: Change in symbol quality represents symbol change in the manuscripts. Repetition of symbols represents same symbol quality, but different interpretation for word (as in Example 9, Figure 5)
% - Lack of recognition of position by one or more processors
-- Utterance deemed indistinguishable by one or more processors

Figure 7. Attested Choice Sets for All Versions of Set I

Sets of Choices	Attested Symbols (%)	(--)	(,)	(..)	(..#)	(?)	(.)	(!)	No. of Options	Total Positions	Frequency (Potential Positions) I E91	II E94
1.	%	--	,				.		(4)	1	1	0
2.	%	--	,						(3)	5	2	3
3.	%	--	,	..			.		(5)	2	2	0
4.	%	--	,	..					(4)	2	2	0
5.	%	--		..			.		(4)	1	1	0
6.	%	--		..					(3)	2	2	0
7.	%		,	..		?	.		(5)	1	0	1
8.	%		,	..		?			(4)	1	1	0
9.	%		,	..			.		(4)	4	2	2
10.	%		,	..					(3)	9	2	7
11.	%		,		..#				(3)	1	0	1
12.	%		,			?	.		(4)	1	0	1
13.	%		,			?			(3)	4	3	1
14.	%		,				.	!	(4)	1	1	0
15.	%		,				.		(3)	10	0	10
16.	%		,					!	(3)	1	1	0
17.	%		,						(2)	96	35	61
18.	%			..	..#	?			(4)	2	0	2
19.	%			..	..#				(3)	1	0	1
20.	%			..			.		(3)	2	0	2
21.	%			..					(2)	1	1	0
22.	%				..#		.		(3)	2	0	2
23.	%				..#				(2)	1	0	1
24.	%					?	.		(3)	8	5	3
25.	%					?			(2)	6	3	3
26.	%						.		(2)	42	19	23
27.			,	..		?			(3)	4	3	1
28.			,	..			.		(3)	2	2	0
29.			,			?			(2)	4	3	1
30.			,				.		(2)	25	7	18
31.						?	.		(2)	32	24	8
32.				..			.		(2)	1	0	1
33.			,						(1)	1	0	1
34.						?			(1)	16	11	5
35.							.		(1)	48	23	25
E91 + E94												
35	26	6	20	15	5	11	17	2	(8)	102	340	156	184

Key:

% - Lack of recognition of any symbol for a position for which one or more processors had indicated a choice.

-- Indication of an utterance the character of which was indistinguishable by one or more processors.

Figure 8. Distribution of Attested Symbol Choices Made by One or More Processors of Sets I and II for All Attested Potential Punctuation Positions.

permutation sets, or how to differentiate the 'zero' of 'indistinguishable' sets from permutations which have definite punctuation symbols for each choice, is another problem. For some analyses the unanimous indication of definite punctuation position (especially if it can be determined whether the indication represents an acoustic or a semantic feature of separation) will be significant, even if there is not agreement on the specific symbol. For others, perhaps all occurrences of % or -- will be as significant as definite symbols.

Note, by the way, that all permutations of the 8 attested symbols do not exist for the manuscripts under examination.

Figure 9, for instance, indicates that there was only one five-way choice for each set, but that the choice was different for each of them (See Figure 8, sets 3 and 7). Figure 10 gives the attested permutations for the illustrative raw text samples of Figure 2. These are fewer, of course, than those attested for all versions of Set I. Note in Figure 11 that the choice for the final editor did not always represent the majority choice. Note also that the most frequent membership for the choice sets was two or three options.

Figures in Fig. 6, therefore, are formally realistic for Initial Universal agreement only. Multiple-categoried discrepancies confuse the absolute statistics for both terminals and initials. Even if the total number of potential initials were to equal the total number of potential terminals (usually sentence-initiating capital is redundant with a preceding double space which in turn is sometimes preceded by a terminal signalling end of preceding sentence, and therefore mandatory initiation for the next symbol), there would be no assurance from the figures that there would, in fact, be that number, or fewer, or more initials or terminals in the 'accurate' or canonic version (here represented tentatively as I,G (N)).

Tentative predictions for terminal representation then, is that there will be enough disagreement to be potentially significant for type and duration (as well as quality) of utterance. This, combined with the frequency of discrepancies for word and phrase omission (Hays 1968a,b, 1969b, 1970a,b) indicates that at least several iterations of verifications of the transcription against the tape by processors equipped with explicit and consistent directions will be necessary to represent utterance boundaries with sufficient reliability to be predicted, recognized and used by an interdisciplinary group of analysts working on or interpreting the same material.

Conclusions

A sample of two transcriptions is obviously much too small to make definitive judgments about oral-graphic language structures. More meaningful statements might be made with a larger sample supported by statistics on the graphic usage (perhaps from a survey of punctuation in college essays, works of literature and everyday writing styles) as well as on the occurrence of speech pausal phenomena (as in Feldstein's work, for example). Controlled experiments on transcription punctuation practices are projected by me for the future, and perhaps will throw

Transcription Sets	Total Option Sets	Number of Options and Frequency of Attestation										Total Positions
		(5)	Freq.	(4)	Freq.	(3)	Freq.	(2)	Freq.	(1)	Freq.	
I	24	1	2	6	8	8	20	7	92	3	3	325
II	25	1	1	3	5	10	3	8	116	2	3	?
I + II	35	2	3	8	13	13	5	9	208	3	6	340

Figure 9. Distribution of Membership Size for Punctuation Symbol Option Sets Attested for Sets I and II

Sentence-Initial		Internal		External		
Upper/Lower Case	Upper/ Upper	,/%	,/%/...	./?	?/---	%/.
2	2	10	2	3	1	1

Figure 10. Frequency of Types of Punctuation Discrepancies Attested For the First Few Lines of Text of All Versions, Raw and Edited, of Set I, Illustrated in Figure 2

	N's Choice	Potential Positions For Symbol	Total Attested Symbol Options	Frequency of Choice per Symbol For all Processors (%)	(--)	(,)	(..)	(..#)	(?)	(.)	(!)
1.	%	83	7	26	1	30	3		5	18	0
2.	--	10	5	5	1	1	1			2	
3.	,	66	7	17	0	39	2		2	6	0
4.	..	25	6	3	1	8	3		8	2	
5.	..#	0	0								
6.	?	52	5	3	6	0			38	5	
7.	.	87	7	8	0	6	1		23	49	0
8.	!	2	4	0		1				1	0
Totals:	8	325	7	62	9	85	10	0	76	83	0

(Note: The choice of Set I-G's Final Editor (N) did not always agree with the choice of the majority of processors for the full set of versions for I)

Figure 11. Frequency of Punctuation Symbol Choice for Final Processor (N) of I-G Against Choices for All Processors of Set I.

more light on the subject.

In the meantime, this small sample can establish that punctuation of oral utterances translated into graphic form is not always consistent, even among variously well-educated processors.

This seems to place in question the inherent reality of the sentence as an inviolable unit of speech. If analysis of speech is based on the sentence units represented in the traditional orthographic manner, it is conceivable that alternate results would be produced for alternant versions of the same stretch. Grammatical analyses of oral speech based upon specific graphic, sentence-bounded structures may be difficult to apply legitimately to natural discourse. In view of the apparent arbitrariness with which punctuation marks are assigned, it may be that analysts of oral empirical evidence will have to reformulate their criteria and choose a unit more resembling the phrase than the sentence as the basis of grammatical discussion.

Research tied to oral sentence types: declarative, imperative, question, etc., may have to be re-examined in this light, especially if these forms eschew so-called 'incomplete' sentences and thoughts.

It is, of course, possible that the difference between speech and writing is only in surface structure and would differ simply in the apportionment of traditionally viewed embedded sentences or sentence partials. But it also might be that there is a deeper difference between the phraseological boundaries of speech and those of writing, which may be masked now by the traditional method of converting the speech code to the graphic code, a process of translation which could automatically distort the realities of both systems. It seems obvious that the underlying system on which each is based must be the same American English language, however it be described or explained. Whether there should be different planes of Less-Deep Structure for each of them seems to be worth considering (Hays 1970e).

Certainly for representing in one manifestation the content of the other, an auxiliary system of explanation should be provided. This might serve not only to disambiguate utterances which are similar when taken out of context in one or both systems, but to better transfer the intended or received meaning of the symbols. Applications of these explanatory analyses would be manifold, and especially useful for linguistic and social science researchers investigating empirical evidence of both oral and graphic language structure and behavior.

REFERENCES

Doris, Lillian, and Bease May Miller. 1960 Complete Secretary's Handbook, Revised Edition. Englewood Cliffs, N.J.: Prentice-Hall, Inc., Chapter 17, pp. 300-314.

Feldstein, Stanley. In press. "Temporal Patterns of Dialogue: Basic Research and Reconsiderations." In Studies in Dyadic Communication: Proceedings of a Research Conference on Interview Behavior, edited

by W.A. Siegman and B. Pope. New York: Pergamon.

Hays, Daniel, and Harriett Hays. 1969 Post-Editing Procedures: A Working Manual. Columbia, Missouri: Center for Research in Social Behavior, University of Missouri.

Hays, Harriett. 1968a. "The Need for Post-editing and Standardizing Transcriptions of Classroom Tapes." Memorandum, CRSB, University of Missouri, Columbia, Missouri.

_____ 1968b. "Transcriber Inconsistency Patterns: Study I." Unpubl. Ms., CRSB, University of Missouri, Columbia, Missouri.

_____ 1968c. "Transcriber Training Session and Outline for Study II." Unpubl. Ms., CRSB, University of Missouri, Columbia, Missouri.

_____ 1969a. Transcriber Procedures: A Working Manual. Unpubl. Ms., CRSB, University of Missouri, Columbia, Missouri.

_____ 1969b. "Problems of Converting Speech Events Reliably into a Graphic Form Compatible to Empirical Research in the Social Sciences." Paper delivered to University of Missouri Linguistics Colloquium, December, 1969.

_____ 1970a. "From Mouth to Hand: Obstacles in Rendering Verbal Events Faithfully into Standard Orthography." Paper delivered to the Linguistics Society of America Meeting, July 24. To be included in the ERIC system.

_____ 1970b. "David, Doyle, Darla, Daryl, Dana, Dale, Gill, Keel, Gail, Sheila, or Bill?: Discrepancies among judges for interpretation of the existence and identification of the source (Emitter) of utterances perceived in a natural speech situation." Paper to be delivered to the American Educational Research Association, February, 1971, and included in the ERIC system.

_____ 1970c. "SMASH: A coding system for manuscript analysis of discrepancies among formal, grammatical and semantic strings covering the same oral stretch." (In preparation.)

_____ 1970d. "Consistent Transcribing and Verification Procedures." (In preparation.)

_____ 1970e. "On non-identities of speech and writing." (In preparation.)

Kurath, Hans. 1939. Handbook of the Linguistic Geography of New England. Providence: Brown University.

Lieberman, Philip. 1965. "On the Acoustic Basis of the Perception of Intonation by Linguists." Word 21.40-54.

Ringaard, Kjell. 1964. "The Phonemes of a Dialectal Area, Perceived

by Phoneticians and by the Speakers Themselves." In Proceedings of the Fifth International Congress of Phonetic Sciences (Munster, 1964), pp. 495-501. Basel and New York: S. Karger.

Webster's Seventh New Collegiate Dictionary. 1966. Springfield, Mass: G. and C. Merriam Co.