AN ANALYSIS OF TRANSCRIPTION CONSISTENCY IN SPONTANEOUS SPEECH FROM THE BUCKEYE CORPUS

William D. Raymond[1], *Mark Pitt*[2], *Keith Johnson*[1], *Elizabeth Hume*[1], *Matthew Makashay*[1], *Robin Dautricourt*[1], *and Craig Hilts*[1]

Department of Linguistics[1] and Department of Psychology[2],
The Ohio State University, Columbus
raymond@ling.ohio-state.edu

ABSTRACT

We present a preliminary analysis of transcriber consistency in labeling and segmentation of words and phones in the Buckeye corpus of spontaneous, informal speech. We find that pairwise inter-transcriber agreement on exact phone label match was 76%, and segmentation agreement within 20% of phone pair length was 75%, though longer phones are more consistently segmented than shorter phones. Patterns of consistency variation in labeling are observed as a function of phonetic categories that are similar to patterns reported for read speech. More agreement is seen on consonants than on vowels, and on fricatives and labials than on other consonant classes. In general, we find that shorter, more reduced words and phones result in more transcriber disagreement.

1. INTRODUCTION

Most studies of transcription consistency in labeling and segmentation of speech have examined inter-transcriber agreement on read speech (Eisen, Tillman, & Draxler [2]; Eisen [1], Wesenick & Kipp [4]). Less is known about the degree of consistency and the range of variation that can be expected in the transcription of spontaneous speech. This study measures inter-transcriber agreement in a corpus of spontaneous English speech, the Buckeye corpus. The corpus was collected, and is being transcribed, at the Ohio State University.

We examine the consistency of the lexical and phonetic labels and their alignment produced by transcribers working on the Buckeye corpus. We evaluate inter-transcriber consistency overall, but also consider how consistency varies with some intrinsic characteristics of the phonetic label set and of the phones of English, especially as produced in the informal style of the speech. An additional consideration in the transcription of spontaneous speech is the degree of agreement on word identification and what factors affect identification consistency. Because the Buckeye corpus is intended to form a basis for the investigation of pronunciation variation in spontaneous speech, it is also important to understand the ways in which the transcribers' choices may contribute to the variation exhibited in the transcribed corpus. Finally, we wish to assess the extent to which individual transcriber differences contribute to transcription disagreements.

2. METHODOLOGY

The Buckeye corpus consists of 300,000 words of speech from recorded interviews with forty speakers. The speakers, natives of Columbus, Ohio, are stratified for age and gender. The corpus has been orthographically transcribed using English orthography (less punctuation).

Phonetic segmentation and labeling of the corpus is being carried out in two phases. Corpus speech is first automatically labeled and aligned using *Entropic Aligner* software. The transcription process is completed by manually correcting automatic labels and their placements using the *Entropic x_waves* interface. Transcribers may use the audio speech signal, the speech wave, and spectrograms in transcription of the digitalized speech. In addition to labeling and aligning lexical items and phones, non-speech events (e.g., silences and noise) and dysfluent events (cutoffs, errors, and fillers) are identified. At this time approximately 15% (45,000 words) of the corpus has been phonetically transcribed.

The phone label set used for transcription is the DARPA-based set used by the *Aligner* interface (Wightman & Talkin [5]), supplemented with four additional segment labels, bringing the number of labels to 50. The added symbols are syllabic nasals, a rounded reduced vowel, and a glottal stop.

Consistency in word and phone transcription was assessed by having the four current transcribers of the Buckeye corpus transcribe the same short speech sample. Three of the transcribers had been transcribing corpus speech for at least one year and had discussed and developed transcription conventions together. The project conventions are documented in the Buckeye coding manual (Kiesling & Raymond [3]). The fourth transcriber was new to the project, but is a highly experienced transcriber using *x_waves*. He familiarized himself with the project conventions from the manual and discussions with other transcribers.

The speech sample selected for the preliminary agreement study consisted of one minute of speech starting approximately 10 minutes into the interview with one female speaker over 40. The speech of this participant had not previously been phonetically transcribed. The speech sample consisted of about 200 words and 450 phones. Transcribers worked independently on the sample, starting with the extant English text transcription.

Transcriber agreement was assessed by comparing transcriber labels and time stamps for words and phones pairwise for all six transcriber pairs. Summing agreement by

event pairs (word pairs or phone pairs) for each transcriber pair gives a measure of *inter-transcriber* consistency for words and phones. Comparisons were also made of the word and phone labels themselves, providing measures of *label agreement*, and of label alignment with the sound wave, providing measures of *segmentation agreement*. *Transcription agreement* is defined as the combined measures of label and segmentation agreement.

3. RESULTS

3.1. Word label and word segmentation agreement

Unlike in the labeling of read speech, identification of words in spontaneous speech allows for uncertainty that may occasionally result in the selection of different words by different transcribers. Nevertheless, word label agreement was high. Transcribers agreed on word identity in 98% of the transcriber pairs (N=882 pairs). However, we would like to know under what circumstances transcribers are likely to disagree on word identification.

In the test sample, word label disagreements occurred where one transcriber identified a word at a point where another transcriber did not, or where transcribers assigned different labels to words. Words on which transcribers disagreed were shorter on average than other words. Disagreements averaged 114 msec in duration, while the average length of all words in the sample was 315 msec. Words differentially labeled also tended to contain sequences of segments plausible for either labeling. For example, the phone label sequence [ɪs] was chosen by two transcribers to label a sequence of two segments; however, the sequence was labeled with the words *I was* by one transcriber and simply as *was* by the other. The word sequence *and all* was identified by three transcribers, with corresponding phones labeled [ænO] (O=open o), [æñO] (ñ=nasal flap), and [ña]; the fourth transcriber labeled the stretch of speech *you know*, phonetically labeled [ɛnO]. The events that were labeled differently by transcribers all involved function words or discourse markers that were short and substantially reduced in form as a consequence of the spontaneous style of the sample.

Word segmentation agreement was measured by comparing placement of word labels across transcribers. When segmentation equivalence for word placement is required to be exact (i.e., less than 1 msec), transcribers agreed on 43% of the transcriber pairs. As would be expected, word segmentation agreement increases as agreement tolerance increases. Mean segmentation placement difference across all transcriber pairs was 26 msec, at which tolerance agreement was nearly 90%.

3.2. Phone label agreement

Phone transcription agreement can be most simply measured as exact phone label match within the set of 50 available phone symbols. Transcribers agreed on phone identity in 76% of all transcriber pairs labeling the same phone (N=2624 pairs). This means that, on average, three or four of the four transcribers agreed on the identity of each phone. Labeling consistency is thus high given the stringent nature of exact label matching.

In order to explore how phone characteristics affect transcriber consistency, we considered label agreement within various classes of phones across different dimensions. Agreement in these classes is measured by calculating percentage of segments on which all transcribers agree. Note that unanimous agreement is an even more stringent measure of consistency than pairwise agreement.

Transcribers all agreed on the existence of a phone (though not necessarily on its label, that is) for 86% of the events identified as a phone by at least one transcriber (N=487 events). It is instructive to consider the segment identification differences, because they result in transcription differences that result in different word variants of the same words. Of the 43 disagreements, 18 are disagreements over final segment deletion (e.g., [fan] v. [faɪnd] for *find*), 13 over non-final deletion ([hæz] v. [æz] for *has*), and 6 involve syllabic segments (*people* with [l] or syllable [l]). The remainder identification differences are due to word label disagreements.

Segments may be divided into the broad categories consonant and vowel. Across consonants (N=282 events), transcribers all agreed that the segment was a consonant 82% of the time. The agreement by all transcribers that an event was a vowel was 83% (N=215 events). Further investigations of phone categories consider consonants and vowels separately,

Consonant phone labels were first conflated into five equivalence classes by manner of articulation for consonants (vowel, glide, stop, fricative/affricate, and nasal/liquid). Consonants were identified as having the same manner by all transcribers 74% of the time. Figure 1 shows that agreement varied as a function of manner. Glides have the lowest consistency, and stops show more disagreement than fricatives/affricates, as has been reported in other studies ([1], [4]).

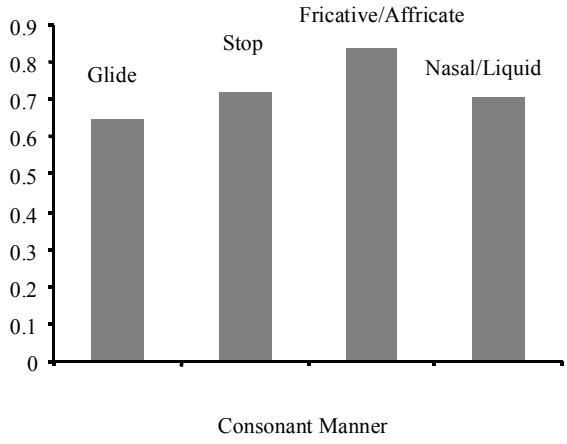


Figure 1: Consonant label agreement by manner class.

We next asked how consonant label agreement varied by the articulatory classification of place. Consonant labels were conflated into five place classes (labial, alveolar, post-alveolar, velar, or glottal). Transcribers gave consonants labels indicating the same place of articulation 71% of the time. Figure 2 reveals some differences in labeling consistency among place categories. Labial agreement was highest. Glottal labeling was less consistent than most other classes. The lower consistency in glottal labeling parallels results from studies using read speech ([1]).

Vowels were first categorized as reduced, diphthong, and unreduced monophthong. Figure 3 shows that monophthongs and diphthongs had comparable label agreement, but there was essentially no agreement on vowel identity in the reduced vowel class. The low agreement among reduced vowels is a product of their high confusability with each other and with other vowels, (see §3.3).

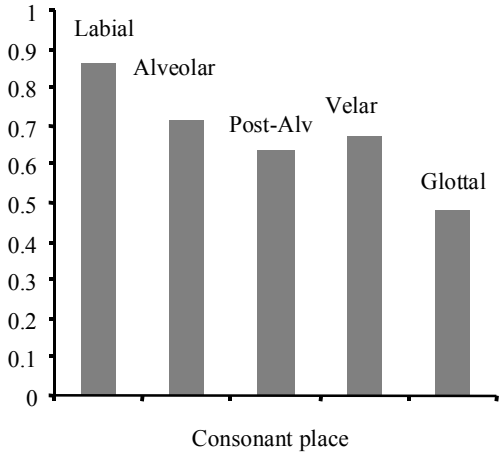


Figure 2: Consonant label agreement by place of articulation.

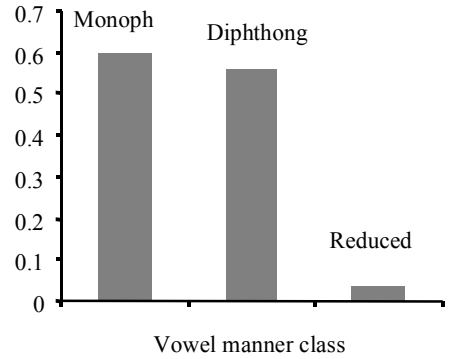
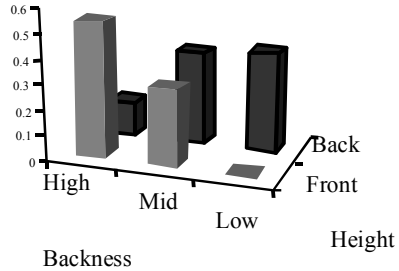


Figure 3: Vowel label agreement by manner class.
Figure 4: Vowel label agreement by height and backness.

Overall, vowel agreement was only 44% across the three categories; even excluding the reduced vowels agreement was only 58%. Agreement on vowels even broadly categorized is thus lower than for consonants categorized by place or manner.



Vowels (excluding diphthongs) were also conflated into place classes in two dimensions by backness (front, back) and by height (high, mid, low). Transcribers labeled vowel backness the same on 52% of the vowels, and labeled vowel height the same on 49% of the vowels. Figure 4 shows vowel label consistency in a two-dimensional vowel space. The greatest consistency in labeling is seen for high front and low back vowels. The low front category contains only the vowel [æ] and reflects the low agreement on this vowel.

3.3. Phone confusion

We have looked at the distribution of disagreements over a number of categories, but have not examined what alternatives transcribers select when they disagree. Tables 2 and 3 show consonant confusion by transcriber pairs for consonant manner and place. The most frequent confusions on consonant manner included differential categorization of stops as fricatives/affricates (e.g., [trai] vs. [chrai] for 'try') or as nasal/liquids (including flap v. nasal flap). Confusion between places mainly involve categorization of underlying *t*'s as alveolar ([t],[d], flap, or nasal flap), as post-alveolar (e.g., the 'try' case above), or as a glottal stop. Glides show little interaction with other categories, but are frequently confused with vowels (38 transcriber pairs). The confusion of glides with vowels accounts for the low agreement on glides shown in Figure 1.

Table 2: Consonant confusion by manner class.

	Glide	Stop	Fricative/ Affricate	Nasal/ Liquid
Glide	157	2	2	0
Stop		583	15	11
Fricative/ Affricate			441	8
Nasal/Liquid				430

Table 3: Consonant confusion by place class.

	Labial	Alveolar	Post-alveolar	Velar	Glottal
Lab.	329	5	0	0	0
Alv.		798	16	6	29
P.-A.			55	0	0
Vel.				139	0
Glot.					88

Disagreements on vowels are more common than on consonants, and are also show regularities. In 354 transcriber pairs transcribers disagreed on vowel labels. They can be broken down as follows:

	No. of disagreements
• Within the set of monophthongs	164
• Reduced vowel v. Nonreduced vowel	135
• Within the set of reduced vowels	39
• Monophthong v. Diphthong	16

The most common monophthong label disagreement (31 instances) involved disagreements between [i] and [ɪ]. Most other monophthong disagreements involve either [ɛ] (51 disagreements) or [ʌ] (47 disagreements). For example, 22 times transcribers disagreed over labeling a vowel as [ɛ] or [æ]; 15 times the disagreement was over using [ɛ] or [ɪ]. There was disagreement over [ʌ] or [a] in 15 instances, and over [ʌ] or [o] in 14 instances.

The disagreements over labeling a vowel as reduced or unreduced include 86 disagreements between a lax vowel ([ɪ], [ɛ], or [ʌ]) and a reduced vowel. An additional 21 of the disagreements in this category were over labeling a rhotic vowel as reduced or unreduced. When transcribers agree that a vowel is reduced, they disagree on the quality 54% of the time.

In general, diphthongs are more consistently labeled than simple vowels, and unreduced vowels are more consistently labeled than reduced vowels.

3.4. Phone segmentation agreement

Phone segmentation agreement was measured by comparing phone time stamps for equivalent segments (N=2813). For agreement within 10 ms, consistency was 62% of transcriber pairs. When placement agreement is relaxed to within 20 ms, consistency increases to 79%. Mean segmentation placement difference across transcriber pairs was 17 ms; the mean maximum difference across pairs was 31 ms.

Segmentation agreement was also measured relative to the length of the phone being labeled, and relative to the length of the two phones that a label placement partitions. When segmentation agreement was measured as placement within 20% of the average phone length, agreement was 60% of transcriber pairs. When segmentation agreement was measured as placement within 20% of the average length of the two events partitioned by the label, transcribers agreed on 75% of all pairs. Agreement proportional to phone length (within 20%) was greater on phones of greater than average length (73% of pairs) than on phones less than average length (50% of pairs), indicating less consistency on segmentation of shorter phones.

3.5. Inter-transcriber consistency

Whether the results reported above are general characteristics of the transcription process or are exaggerated by inter-transcriber differences can be assessed by considering agreement by transcribers.

Table 4: Label and segmentation agreement for the four transcribers (percentages of transcriber pairs).

Transcriber	Label agreement	Segmentation agreement
A	73%	62%
B	75%	64%
C	76%	64%
D	75%	60%

Table 4 shows label and segmentation agreement for all transcribers. Agreement percentages confirm that the transcribers were similar in their agreement with other transcribers on label choice and label placement. Note also that transcriber A, who was new to the project (although an experienced coder), did not perform differently from the other three transcribers.

4. CONCLUSIONS

We have examined some parameters of consistency in the lexical and phonetic transcription of a sample of spontaneous speech. While word transcription was extremely accurate in general, there was a small amount of disagreement among transcribers regarding the identification of some items. Word disagreements involved function words or discourse markers that were shorter in duration than average and had undergone substantial phonetic reduction in this informal speech. Phone labeling was, as would be expected, less consistent than word identification, but in addition there were differences in the consistency of phone labeling by phone class. Glides and stops were less consistently labeled, and fricatives and labials more consistently labeled, than other classes. The observed consonant differences are consistent with manner and place disagreement differences reported for read speech. Vowels were generally less consistently labeled than consonants in the sample, though again consistency varied by vowel type, with greatest consistency seen for unreduced vowels, especially those that are phonetically complex (the diphthongs) and those farthest from the reduced vowels in the vowel space (high front vowels). Given the differences in transcription consistency between words and phones, and within word and phone classes, it seems reasonable to conclude that one factor affecting transcriber agreement, and hence the reliability of transcription, is the size of the transcribed unit, with smaller units showing less transcription agreement. Unsurprisingly, the greater uncertainty of the spontaneous, informal speech translated into somewhat more disagreement in its transcription than reported for read speech. However, the patterns of disagreement reflected the intrinsic phonetic factors previously identified as correlated with transcription inconsistencies, along with some additional phonetic and lexical factors that are a consequence of the informal style of the Buckeye corpus.

5. REFERENCES

[1] Eisen, B., Reliability of speech segmentation and labeling at different levels of transcription. *Proceedings of Eurospeech* 1991, Berlin/Germany, 673-676.

[2] Eisen, B. and H. G. Tillman., Consistency of judgments in manual labeling of phonetic segments: The distinction between clear and unclear cases. *Proceedings of ICSLP* 1992, Banff/Canada, 871-874

[3] Kiesling, S. and W. D. Raymond, The ViC transcriber's manual: Guidelines for transferring, transcribing, and labeling sound files for the Buckeye corpus. MS, the Ohio State University. 2000.

[4] Wesenick, M.-B. and A. Kipp, Estimating the quality of phonetic transcriptions and segmentations of speech signals. *Proceedings of ICSLP* 1996, Philadelphia/USA, 129-132.

[5] Wightman, C. and D. Talkin, The Aligner user's guide. Entropic Research Laboratory, Inc. Washington D.C./USA. 1997.