An analysis of coding consistency in the transcription of spontaneous speech from the *Buckeye* corpus

William D. Raymond
Ohio State University

1. Introduction

Large corpora of speech that have been supplemented with linguistic annotation, in the form of aligned category labels at a variety of linguistic levels, have proved an important tool for phonetic and phonological analysis (Jurafsky, Bell, Gregory, & Raymond 2001; Greenberg, Hollenback, & Ellis 1996; Keating, Byrd, Flemming, & Todaka 1994; *inter alia*), and speech technology research (Byrne, Finke, Khudanpur, McDonnough, Nock, Riley, Saraçlar, Wooters, & Zavaliagkos 1998; Deng 1998; Hayakawa, Kato, Sagisaka, & Shirai 2002; Strik & Cucchiarini 1996). In using linguistically annotated corpora, the degree of consistency in labeling and alignment and the reliability of the resultant transcriptions have important implications for the value of the corpus as data. However, manual transcription is a subjective process, and one that is prone to error.

Studies of transcription consistency in labeling and alignment of speech have examined inter-transcriber agreement on manual alignment of read speech (Eisen 1991; Eisen, Tillman, & Draxler 1992; Wesenick & Kipp 1996) and conversational speech (Greenberg 1998), and even agreement of automatically aligned speech with manual transcription (Saraçlar & Khudanpur 2000). Transcription agreement on read speech among expert human transcribers can be high, though this does not guarantee transcription accuracy (Cucchiarini 1996). Wesenick & Kipp (1996) found 94.8% label agreement and 93% alignment agreement within 15ms for a corpus of read German sentences. In a study of transcription of the *Switchboard* corpus of conversational English speech, Greenberg (1998) reports inter-labeler agreement in 75-80% of transcriber pairs, considerably lower than agreement on read speech. Interestingly, Saraçlar & Khudanpur (2000) found that for the *Switchboard* data used in Greenberg (1998), alignment agreement of automatically aligned speech with hand transcription was 75%, comparable to inter-labeler agreement on this speech.

We can conclude that regardless of the speech source or transcription method, transcribers are not consistent in label selection and alignment. However, the variation among transcribers is not without patterns. Previous studies of transcription consistency have identified a variety of factors associated with how transcriptions vary across transcribers when they are annotating the same speech. In addition to speech style and speaker dialect (Kerswill & Wright 1991), phonetic transcription variability has been shown to be sensitive to phonological features and context (Eisen 1991; Eisen, Tillman, & Draxler 1992; Wesenick & Kipp 1996), and also lexical context (Shriberg & Lof 1991). Transcription consistency is thus a function of some of the same factors that influence pronunciation variability. Consequently, we might expect other factors influencing pronunciation variation to be associated with lower inter-transcriber consistency, such as metrical and prosodic context (Greenberg 1998) and unit probabilities (Jurafsky et al. 2001). It is not surprising, then, that transcription of spontaneous speech shows less consistency than transcription of read speech, even when transcription is done by trained transcribers, because spontaneous speech has been found to result in a broader range of pronunciations than read speech (Keating 1997). The lower consistency in transcribing spontaneous speech underscores the importance of assessing where disagreement is likely to occur, both as a guide to reliability of the transcription in use of data from a corpus and as a means of improving consistency of future corpus transcription.

The current study was conducted to measure inter-transcriber consistency in labeling and alignment of a corpus of spontaneous English speech, the *Buckeye* corpus. The corpus was collected, and is being linguistically annotated, at the Ohio State University. The study examined the consistency with which transcribers working on the *Buckeye* corpus applied lexical and phonetic labels standard to the project, and also the consistency of label alignment to the speech signal. The Buckeye corpus is intended to form a basis for the investigation of pronunciation variation in spontaneous speech, so it is important to understand the ways in which the transcribers' choices may contribute to the variation exhibited in the corpus transcription.

Consistency in lexical and phonetic transcriptions was analyzed with respect to a variety of intrinsic characteristics of the English words and phones, especially as produced in the informal style of the corpus. Phonetic factors that have been claimed to influence segment transcription consistency in past studies were analyzed, as well as some additional factors, such as phone frequencies, phone durations, and prosodic prominence. In addition to the phonetic features, the influences on segmental and lexical labeling of some properties of the words containing the transcribed phones were also examined. Because of the spontaneous style of speech in the corpus, the degree of agreement on word identification was also analyzed to determine what factors affect identification consistency of words.

The paper begins with a brief discussion of word labeling and alignment consistency and where transcribers disagree on word identification. The focus of the study, which follows the word level analysis, will be on the consistency of phone labeling and how consistency of phone labeling and alignment vary. Five classes of factors were examined in the study in analyzing the consistency of phonetic transcription: (1) phonological features; (2) phone frequencies; (3) phone durations; (4) lexical properties of the words containing the phones; and (5) the prosodic environment in which the phone is produced. After presenting the results of the analyses, the paper concludes with a review of the results.

2. The *Buckeye* corpus and its transcription

The *Buckeye* corpus consists of over 300,000 words of speech gathered through recorded interviews with forty speakers. The speakers, natives of Columbus, Ohio, were stratified for age and gender. Talkers were asked by express their opinions on a variety of topics, facilitated by an interviewer. The interview recordings have been digitalized and orthographically transcribed using standard English orthography (less punctuation) and a few special forms representing frequent collocations with special meanings or functions (e.g., *gonna* for *going to* as a future tense marker, and *yknow* for *you know* as discourse filler) to enable recognition of the marked functions of the forms. In the orthographic transcriptions annotations were made of the occurrence of interviewer speech (but not its content), as well as non-speech events, including vocal noise (e.g., coughing or laughter) and background noise. Dysfluencies were also annotated in the orthographic transcription, including filled pauses (*uh* and *um*), lexical cutoffs, and speech errors.

Phonetic labeling and alignment of the corpus is being carried out in two phases. Corpus speech is first automatically labeled and aligned using the *Entropic Aligner* software. The transcription process is completed by manually correcting automatic labels and their placements using the Entropics product ESPS/*waves+*. Transcribers use the audio speech signal, the speech wave, and spectrograms in labeling and alignment of the digitalized speech. In addition to labeling and aligning lexical items and their phones in the talkers' speech, dysfluent events (lexical cutoffs, speech errors, and fillers) are also labeled and phonetically aligned. Non-speech events (silences and noise) and interviewer speech are labeled and aligned, although as in the orthographic transcriptions interviewer speech is not phonetically transcribed. At this time automatic alignment has been completed on about 60% of the corpus, and manual correction has been completed for 30% of the corpus from 13 speakers, representing over 100,000 words of speech.

The goal of corpus transcription has been to represent the speech as discrete segments that capture pronunciation variation, including some allophonic variation, and that are consistently applied. However, the level of detail represented by the label set was different for different segments, depending on project-specific interests. For research requiring greater or lesser transcription detail, labels are intended to serve as indices into the speech signal, where additional acoustic and auditory information may be accessed. Labels chosen for transcription reflect these goals. The phone label set used for transcription of the corpus is the DARPA-based set used by the *Aligner* interface (Wightman & Talkin 1997), supplemented with some additional segment labels. The added symbols are nasal vowels (one for each oral vowel), syllabic nasals, a rounded reduced vowel, and a glottal stop. With the supplemental labels, the number of symbols in the label set is 69.

3. Methodology

Consistency in word and phone transcription was assessed by having four transcribers of the *Buckeye* corpus label and align four one-minute samples of corpus speech from four different talkers. 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 used by transcribers are documented in the *Buckeye* coding manual (Kiesling & Raymond 2000). The fourth transcriber in the study was new to the project, but was a highly experienced transcriber with extensive *ESPS* software familiarity. He familiarized himself with the project conventions from the manual and discussions with other transcribers before the consistency study.

The speech samples selected for study each consisted of one minute of speech starting approximately 10 minutes into the interviews with four different talkers. The four talkers selected for the study were balanced for age (over and under 35) and gender in a 2X2 design. The speech samples selected had not previously been phonetically transcribed. The samples consisted of about 730 words (including dictionary words and other types of "word events", such as lexical cutoffs and fillers) and about 2300 phonetic segments. Transcribers worked independently on the samples, starting with the extant English text transcriptions for each sample. In addition to word and phone transcription, one transcriber, trained in the ToBI system of prosodic annotation (Beckman & Hirschberg 1994, Pitrelli, Beckman, & Hirschberg 1994), provided prosodic coding of pitch accent placements and prosodic boundaries for the samples.

When the transcriptions had been completed, the labels were equated for comparison across transcribers by the author. Transcriber agreement was assessed by comparing the transcriber labels and time stamps for equivalent word and phone events pairwise for all six possible transcriber pairs. Summing disagreement pairs for each event (word event pairs or phone event pairs) for all transcriber pairs and dividing by the total number of pairs provides the proportion of disagreements across transcriber pairs. Disagreement proportions were calculated for word and phone labels, or groups of labels, providing measures of *label agreement*, and for label alignment with the sound wave, providing measures of *alignment agreement*.

In examining transcriber consistency we would like to know not only the degree to which transcribers agree on label choice and label placement, but also where transcribers are likely to disagree and what factors influence consistency. Because transcribers labeled and aligned both word events and phone events, we considered consistency for both transcription levels.

4. Word transcription agreement

There were 734 words in the study sample. Transcribers collectively identified 2806 word events, creating 4072 transcriber pairs of word events. As would be expected, word label agreement was high. Transcribers agreed on word identity in 99% of the transcriber pairs.

Unlike in the labeling of read speech, identification of words in spontaneous speech allowed for some uncertainty that resulted in the selection of different words by different transcribers, or even the identification of a word by one transcriber with no corresponding word in another transcriber's word labels. In the test sample there were 22 word events with labeling disagreements (10 word label disagreements and 12 word identification disagreements), totaling 74 disagreement pairs for word events.

Word label disagreements tended to contain sequences of segments plausible for either labeling. For example, the words at the end of the phrase in (1) were transcribed with four different sets of phone labels and two different word label sequences. The word sequence *and all* was identified by three of the four transcribers, with labels shown in (1). The fourth transcriber phonetically labeled the stretch of speech [ɛnɔ], which could also plausibly be a transcription for *and all*; however, the transcriber labeled the speech with the words *you know*. Speech context could not be used to decide between the two word transcriptions in this case, because both sequences were plausible in the semantically and syntactically unconstrained context.

(1) *people enjoy talking with people* <pause> *and getting out and* <pause> __________

Transcriber	Phone transcription	Word labels
1	[ænɔ]	*and all*
2	[æ̃ɾ̃ɔ]	*and all*
3	[ɾ̃a]	*and all*
4	[ɛnɔ]	*you know*

As in (1), word label disagreements largely involved function words (especially pronouns and prepositions) or frequent collocations (e.g., *you know*, *in front of*) that were substantially phonetically reduced in form as a consequence of the spontaneous style of the sample. Word label confusions were generally between words of the same type, for example, both pronouns (*them* and *him*) or articles (*a* and *the*). The words which transcribers labeled differently were also shorter on average than other words. Words with different labels averaged 154 msec in duration, while the average length of all words in the sample was 240 msec. Word label disagreements occurred on events that were even shorter than most function words, which averaged 176 msec.

Word identification disagreements involved words that were predictable or expected from the context. For example disagreement over identification of *of* in *in front of* (identified by only two transcribers) and disagreement over identification of *I* in *I was* (identified by three transcribers). Word identification disagreements were not as short as word label disagreements (201 msec).

The results indicate that word label disagreements occurred between words that were phonetically and functionally similar, and where contextual information could not be used to differentiate the alternatives. On the other hand, word identification disagreements occurred on contextually predictable words.

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

5. Phone transcription agreement

There were 2305 unique phone events in the speech sample. Transcribers identified 8601 phone events, creating 12,370 transcriber pairs of phone events. Transcribers did not always agree on the existence of a phone in the speech signal. Agreement on phone existence was 92% of the (13,830) possible transcriber pairs. Phone identification disagreements were not further analyzed. Proportions of transcriber pair agreement on phone labeling reported below will be out of the 12,370 pairs in which both transcribers in the pair labeled a phone.

Phone transcription agreement can be most simply measured as exact phone label match within the set of 69 available phone labels. Transcribers agreed on phone label in 79% of all transcriber pairs labeling the same speech segment. This means that, on average, one transcriber disagreed on a label for about every two speech segments labeled in the test sample. Labeling consistency was thus comparable to consistency rates in other examples of transcribed spontaneous speech (Greenberg et al 1996). The proportion of agreement in exact label match for each pair of transcribers was quite consistent, ranging from 79% to 81%, as seen in Table 1, indicating that transcribers were very similar in their labeling judgments. This similarity in consistency for all transcriber pairs was also seen for the other factors examined in the study.

Table 1: Exact phone label match by pairs of transcribers.

Transcriber pair							
	1-2	1-3	1-4	2-3	2-4	3-4	All pairs
Number of labeled pairs	2082	2059	2048	2067	2065	2049	12,370
Proportion exact label agreement							
Speaker 1 (Female, younger)	.75	.73	.76	.78	.78	.76	.76
Speaker 2 (Female, older)	.77	.81	.78	.81	.80	.77	.79
Speaker 3 (Male, older)	.83	.83	.82	.83	.83	.84	.83
Speaker 4 (Male, younger)	.79	.80	.77	.80	.76	.79	.79
Total	.79	.79	.78	.81	.80	.79	.79

Table 1 also reveals that label consistency varied somewhat across the four talkers used in the test sample, suggesting that speaker differences can affect transcription consistency. One talker, speaker 3, was transcribed more consistently than the other talkers. There was a difference in transcription consistency by gender (p=.001) as well, with more agreement by pairs of transcribers on the two male speakers (80%) than on the two female speakers (77%). There was no age difference in consistency of transcription. The talker differences may have been the result of group differences or individual differences, but the sample is too small to distinguish the two possibilities.

Table 2 shows the proportions of transcriber pairs on which transcribers disagreed for all labels selected by transcribers at least 10 times in the test sample, and the phones to which they correspond. Some phone labels are clearly more consistently selected than others. Disagreement rate on phone labeling ranges from 3% of transcriber pairs for the label ‘f’ to 97% for the nasalized vowel ‘ihn’.

The label ‘ihn’ was not the only label to be so idiosyncratically applied. It appears in Table 1 because it was the only nasalized vowel to be selected more than 10 times by transcribers, but other, less frequently used nasalized vowel labels (not shown in Table 2) were also very inconsistently applied. As would be expected, inconsistency on selecting the label for ‘ihn’ accounts for much of the disagreement on its non-nasalized counterpart ‘ih’ (54%), which often alternated with ‘ihn’. Because of their infrequent use and labeling inconsistencies, nasalized vowel labels will not be further considered in the analyses, although we will return to an explanation of their inconsistent selection below.

Table 2: Disagreements within transcriber pairs for all phone labels.

Label	Phone	Prop. Dis.	Label	Phone	Prop. Dis.	Label	Phone	Prop. Dis.	Label	Phone	Prop. Dis.
aa	a	.42	dx	ɾ	.30	ix	ɨ	.84	s	s	.17
ae	æ	.39	eh	ɛ	.47	iy	i	.35	sh	ʃ	.17
ah	ʌ	.53	el	l̩	.63	jh	dʒ	.31	t	t	.27
ao	ɔ	.47	em	m̩	.45	k	k	.03	th	θ	.29
aw	aʊ	.23	en	n̩	.50	l	l	.15	tq	ʔ	.50
ax	ə	.64	er	ɛ̂	.58	m	m	.07	uh	ʊ	.46
axr	ɚ	.66	ey	eI	.20	n	n	.25	uw	u	.23
ay	aɪ	.21	f	f	.03	ng	ŋ	.17	v	v	.13
b	b	.11	g	ɡ	.13	nx	ɾ̃	.58	w	w	.10
ch	tʃ	.30	hh	h	.07	ow	oʊ	.33	y	y	.08
d	d	.23	ih	ɪ	.54	p	p	.07	z	z	.30
dh	ð	.26	ihn	ɪ̃	.97	r	r	.11	zh	ʒ	.77

The average alignment difference for all transcriber pairs was 16.4 msec. As with label selection, alignment placement was consistent across pairs of transcribers, as shown in Table 3. The mean transcriber alignment difference of 16.4 msec for all transcriber pairs was 19% of the mean transcribed segment length of 84 msec. Thus, transcriber agreement on label placement was, on average, within 19% of the segment length.

Table 3: Mean alignment difference (msec) for all transcriber pairs.

Transcriber pair	1-2	1-3	1-4	2-3	2-4	3-4	All pairs
Mean alignment difference (msec)	15	16	17	16	17	17	16.4

5.1. Consistency across segment classes

The effects of inherent articulatory characteristics of segments on transcriber consistency were examined by considering inter-labeler agreement within segment classes defined by syllabicity, consonant manner, and consonant and vowel place of articulation. Label agreement was first considered within the broad categories of syllabic peak (V) and non-peak (C) segment labels. Peak segments include vowels and syllabic consonants. Peak segments function as syllabic nuclei and will be referred to as vocalic segments. Non-peak segments include all other segments and are found in syllable margins. Non-peak segment labels will be referred to as consonant segments. Agreement on CV class was seen in 98% of transcriber pairs, indicating that when two transcribers both identified a phone, they almost always agreed on its syllabic status as peak or non-peak.

Consonant labels were conflated into four equivalence classes by manner of articulation of the corresponding phone: (1) glides ([y], [w], and [r]; (2) stops; (3) fricatives/affricates; and (4) resonants ([m], [n], [ŋ], and [l]). There was little disagreement between pairs of transcribers on consonant manner; consonants were given labels with the same manner by both transcribers in 96% of transcriber pairs. However, disagreement varied by manner class (for consonant labels) and between consonant class and syllabicity.

As shown in Figure 1, the highest rate of consonant manner disagreement per label token was in the resonant class [1]. Disagreements involving a resonant constituted 61% of all manner disagreements. Resonants (and also glides) were most often confused with syllabic peaks. Disagreements on labeling resonants largely involved ‘n’ and another [+nasal] phone label, either ‘en’ (e.g., *didn’t* as [ɾ ɪ n ʔ] or [ɾ ɪ n̩ ʔ]) or ‘nx’ (e.g., *twenty* as [t w ɛ n i] or [t w ɛ ɾ̃ i]). Stops showed more disagreement than fricatives/affricates, as has been reported in studies of read speech (Eisen 1991, Wesenick & Kipp 1996). Stops were most often confused with resonants (e.g., nasal flap and nasal stop, as in *doing all* as [d u ɪ ɾ̃ a l] vs. [d u ɪ n a l]), and stop label disagreements almost always involved homorganic pairs (e.g., ‘b’ or ‘v’, ‘d’ or ‘dh’).

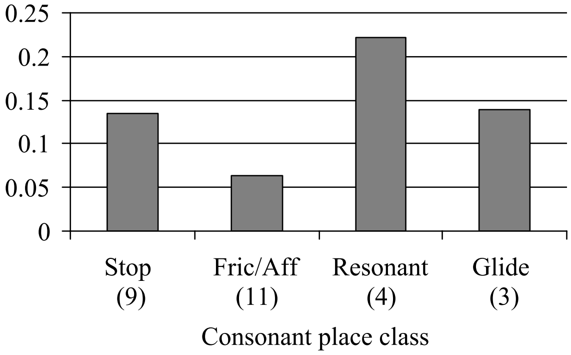

Figure 1: Syllabicity and consonant manner disagreements by class. (Type frequency of each class is shown in parentheses.)

In order to examine place of articulation, consonant labels were conflated into five place classes: (1) labial, (2) alveolar, (3) post-alveolar, (4) velar, and (5) glottal. Transcribers gave C segments labels indicating the same place of articulation 97% of the time, and were thus as consistent in assigning place as they were at selecting consonant manner and syllabicity. Figure 2 reveals some differences in labeling consistency among place categories. Labial place agreement was highest. Glottal labeling was less consistent than most other classes and almost always involved disagreements between labels for [ʔ] and [t]. The lower consistency in glottal labeling paralleled results from studies using read speech (Eisen 1991).

[1] Disagreements in this and subsequent figures are reported as number of transcriber pair disagreements per class token in order to normalize for class size.

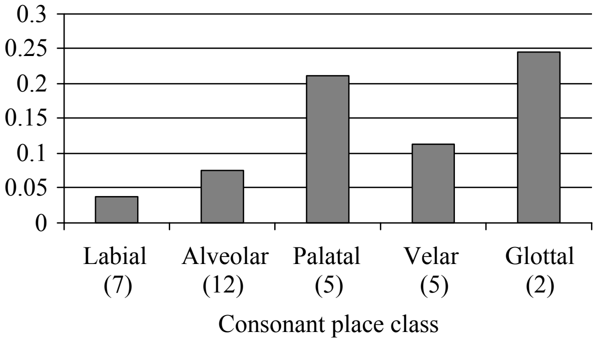

Figure 2: Consonant place disagreements by place class. (Type frequency of each class is shown in parentheses.)

In order to examine the influence of place of articulation on vowel consistency, vowels were categorized as [±high] and [±front], or diphthong. High vowels included the labels ‘iy’, ‘ih’, ‘ix’, ‘uw’, ‘uh’, and ‘ux’; back vowels were ‘uw’, ‘uh’, ‘ux’, ‘ow’, ‘er’, ‘ax’, ‘axr’, ‘aa’, ‘ao’, and ‘ah’; and diphthongs were ‘ey’, ‘ay’, ‘oy’, and ‘aw’. There was agreement on vowel place in 78% of the transcriber pairs labeling vowels, indicated that vowels place was considerably less consistently transcribed than consonant place or manner.

Vowel disagreements were not uniform across place features. Most vowel place disagreements involved front vowels, which accounted for 79% of the vowel place disagreements. Front vowels were frequently confused with diphthongs, back vowels, and other front vowels. A large part of the front vowel disagreement (20%) involved disagreements within the reduced vowel set, especially between ‘ax’ (low back) and ‘ix’ (high front). In fact, the four reduced vowels showed little internal place consistency in transcription. When two transcribers labeled a vowel as reduced, they agreed on the label only 50% of the time. Confusion on reduced vowels usually implies vowel place confusion, because reduced vowels all differ in at least one place dimension. Disagreements of all types within the reduced vowel set accounted for 22% of the vowel place disagreements. With reduced vowel labels removed from consideration, agreement on vowel place features increased to 88% of transcriber pairs, higher than with the inclusion of reduced vowels, but still lower than consonant place consistency.

Figure 3 shows disagreement per class token by vowel place of articulation with the reduced vowels removed. Most disagreements now involved low front vowels and diphthongs. High front vowels were the most consistently labeled. That diphthongs have low labeling consistency as a class means that diphthongs are not confused with other diphthongs; disagreements involving diphthongs are almost always with simple vowel located near one end of the diphthong’s trajectory through vowel space (e.g., ‘ay’ and ‘ih’).

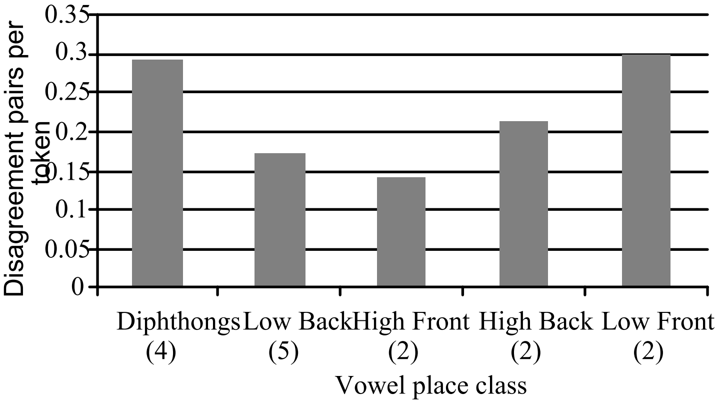

Figure 3: Vowel place disagreements by class (with reduced vowels removed). (Type frequency of each class is shown in parentheses.)

Note that the ordering of vowel categories by consistency showed no correlation with the ordering by number of labels–the type frequency–within a category (shown in parentheses below each category in Figure 3). The same can be said for consonant manner and place classes (see Figures 1 and 2). Thus, the disagreements that occurred are not a consequence of class size. However, in examining inter-class disagreement, label frequency has been controlled. It may be that the rate of disagreement is related to the frequency with which labels are used, a possibility to which we now turn.

5.2. Effects of phone probabilities and phone durations on labeling consistency

Probabilities of linguistic units affect the perception and pronunciation of phones. Higher frequency words and words that are more predictable from their lexical environment exhibit more variation in production than lower frequency words (see Bybee & Hopper 2001). Perception is also facilitated when phones are more predictable (see Pitt & Samuel 1995), as well as when phones are longer. We might expect, then, that unit probabilities, particularly the frequency with which phones occur in speech, may affect the consistency with which phones are transcribed, either by influencing the range of productions of a phoneme or the confusability of its realizations in perception. In addition, because of their greater perceptual salience, longer phones may also be more consistently labeled. To explore the effects of phone probabilities and durations, transcription consistency was examined as a function of the frequency of occurrence of phone labels in the test sample and their average durations. A separate analysis was done for the consonant labels and the vocalic labels.

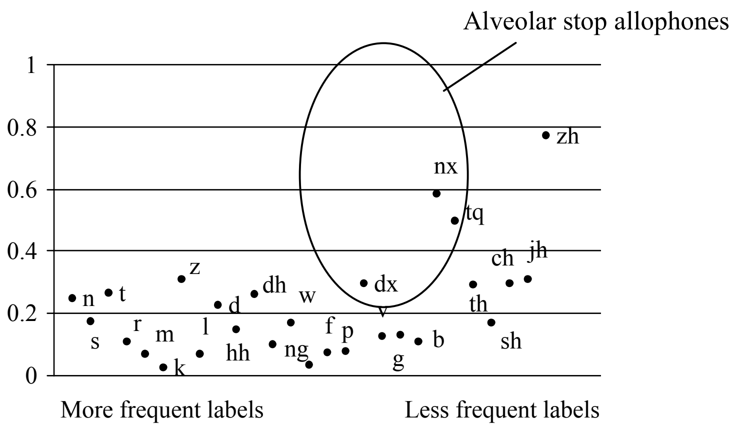

Figure 4: Consonant label disagreements by use frequency.

Figure 4 shows consonant label disagreement as a function of frequency of label use for all labels selected by transcribers. Three outliers on the graph, ‘zh’, ‘nx’, and ‘tq’, are notably different from the distribution of the remaining labels, with much higher disagreement rates than other labels of similar frequency. Two outliers in Figure 4, the nasal flap ‘nx’ and the glottal stop ‘tq’, do not label English phonemes, but allophones of alveolar stops, while all other labels in Figure 4 label phones of English, with the exception of the label for the oral flap ‘dx’. The label ‘dx’, another alveolar stop allophone, also has a high rate of disagreement. While ‘zh’ labels a phoneme in English ([ʒ]), it was selected only 18 times in the sample, fewer than any other label shown, all of which were selected by transcribers in labeling the test sample more than 50 times. In addition, [ʒ] only occurred canonically in one word (*television*), in which labeling agreement on ‘zh’ was 50% of transcriber pairs. All other uses of ‘zh’ labeled non-phonemic variation created through assimilation (e.g., the two occurrences in [ð ɛ ʒ ʒ ɪ s] for *there's just*), and thus did not label transcriptions of canonical pronunciations of underlying phonemes.

After removing the three outliers in Figure 4 there was no appreciable correlation between how often a label was selected and consonant label consistency (r^2 = .042). In addition, there was no strong correlation between consistency of consonant labeling and consonant duration (r^2 = .163).

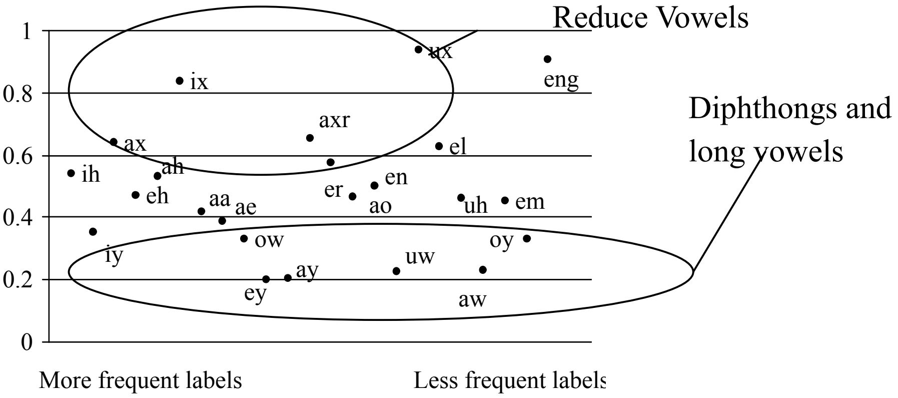

Figure 5: Vocalic label disagreements by use frequency.

The results of the frequency analysis indicate that frequency of consonant label use did not affect labeling consistency. What the labeling consistency of some consonant labels suggests is that label selection was affected by whether the speaker pronounced a phoneme canonically or produced some variant phone as a pronunciation. Because non-phonological labels imply a narrower phonetic transcription, the result is consistent with findings that more narrow transcriptions are less reliable (Shriberg & Lof 1991).

Consistency for vocalic labels as a function of frequency of label use is shown in Figure 5. As with consonant labels, there was little correlation between vocalic label consistency and how often a label was selected ($r^2 = .055$).

What is apparent from the distribution of vocalic labels in Figure 5, however, is a relation between identifiable vocalic subclasses and transcription consistency. Reduced vowels are inconsistently transcribed, as has already been noted. The reduced vowels all showed more disagreement than any other syllabic label except ‘eng’ (used only 4 times). The most consistently transcribed vowels are the diphthongs and the long vowels with off-glides. These six labels all showed less disagreement than any of the other vowels. Unreduced simple vowels and syllabic consonants are intermediate in labeling consistency between the reduced vowels and the diphthongs.

The labeling consistency difference between non-reduced vocalic segments and diphthongs suggests that perhaps label selection is affected by vocalic segment duration or structural complexity. Vocalic segment consistency was not strongly correlated with duration ($r^2 = .11$). Structural complexity is thus the proposed explanation for the relation between non-reduced vocalic labels and consistency.

Finally, because the reduced vowels and ‘eng’ are phonetic variants and not phonemes, high rates of disagreement among reduced vowels may be explained in a way similar to that used to capture disagreement difference among consonant labels: vocalic labels that capture more narrow phonetic distinctions are less consistently applied. Recall that the nasal vowel labels, which are also not English phonemes but allophones (of non-nasal vowels), had high rates of disagreement. For labels of vocalic segments as well as consonant segments, then, transcriptions that make distinctions closer to the phonetic level are less consistent.

5.3. Lexical factors influencing phone transcription consistency

Transcription consistency of phone segments may be influenced by the words in which they are produced. Several properties of words may play a role in transcriber consistency, including word length, word class, and the position of the phone in the word (Shriberg & Lof 1991).

Labels in words transcribed with only a single phone showed substantially more disagreements than labels in words with more than one transcribed phone, as seen in Table 4. However, there is no apparent strong tendency for consistency to increase with word length for words containing more than a single transcribed phone.

Table 4: Label disagreements by word length in transcribed phones.

Word length	Number of words	Total phones in words	Proportion of disagreements
1	76	76	0.49
2	216	432	0.32
3	211	633	0.26
4	112	448	0.29
5	47	235	0.24
6	26	156	0.26
7	18	126	0.30
8	13	104	0.30
9	6	54	0.28
10	1	10	0.15

All the words of length one except one were closed-class words, such as *a* (phonetically labeled 'ax') and *to* (labeled 't'). The one exception was the word *because* transcribed (labeled 'k'). Thus, we cannot tell whether it is length or word class that explains the lower labeling consistency of the singleton words. To explore this issue, the words in the test sample were divided into the two classes of function (closed class) words and content (open class) words. Function words included articles and demonstratives, (personal and indefinite) pronouns, modal and auxiliary verbs, prepositions, and conjunctions. All other words were included in the content word class.

The phones in function words were less consistently labeled overall than the phones in content words. Even when we removed the singleton words, phones in function words were still less consistently labeled than phones in content words. With the singleton words removed, the proportion of transcriber pair disagreements per label occurrence was .36 for functions words and only .22 for content words.

It has been reported that phones at the beginnings of words show less variation than later phones, and greater consistency on initial phones may explain the higher consistency on words of length one. The consistency of initial and non-initial phones differed, with more disagreement on non-initial phones (22%) than on initial phones (17%). The difference was even greater when the singleton words were removed, as shown on the right in Figure 6. There was also more disagreement on non-initial phones in both function words and content words, and the effect was greater for content words than for function words, also shown in Figure 6, again with singleton words removed. Word length did not have an effect on phone labeling consistency beyond the effect of the initial phones of words, even for content words, as shown in Figure 7.

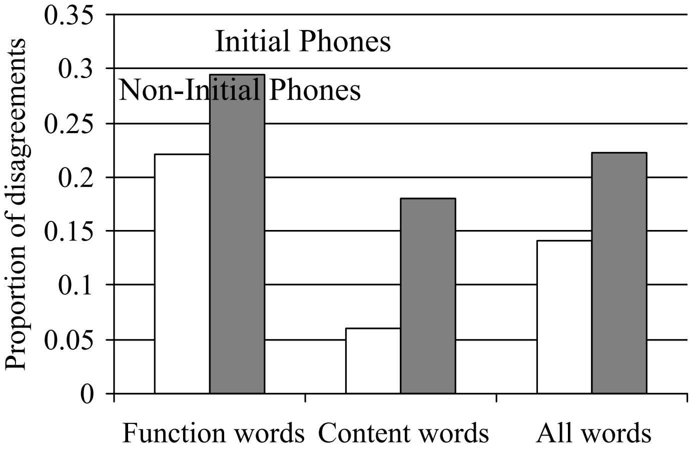

Figure 6: Disagreement in word-initial and non-initial phones.

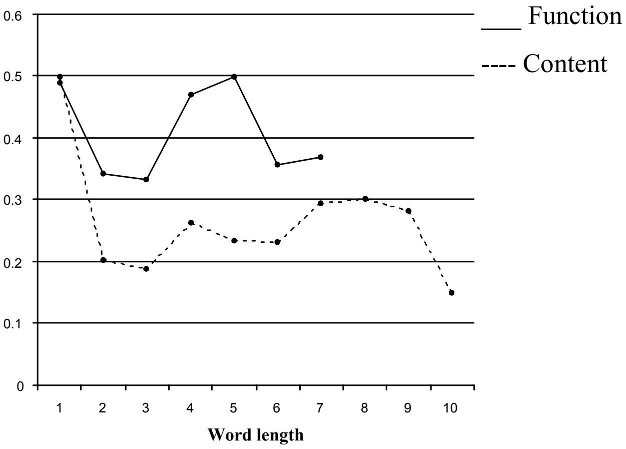

Figure 7: Disagreement by word length and word class.

5.4. The effect of phrasal accent

Phrasally prominent syllables are longer, less reduced, and more acoustically salient than non-prominent syllables. Phrasal prominence involves prosodic marking by change of pitch on one or more lexically stressed syllables in a prosodic phrase. The marked syllables convey the pragmatic function of the phrase. As a result of its acoustic and articulatory effects, phrasal prominence, or its absence, may influence transcription consistency. In particular, we would expect phones in phrasally accented syllables to be more consistently transcribed. To test this possibility, vowel consistency was compared for vowels in phrasally accented and unaccented syllables.

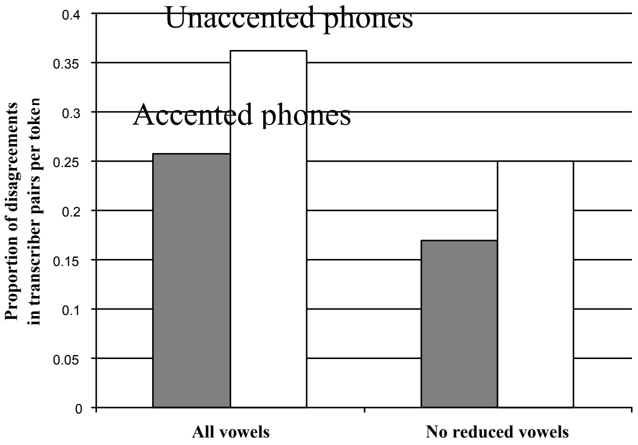

Figure 8: Disagreement by word length and word class.

Vowels in phrasally accented syllables were indeed more consistently transcribed than vowels in non-accented syllables, shown on the left in Figure 8. As we have seen, reduced vowels, which can only occur in unstressed syllables in English (and thus cannot be in a syllable on which a phrasal accent is placed), were very inconsistently transcribed, and so the inclusion of the class of reduced vowels in the prominence comparison may account for the lower labeling consistency of unaccented vowels. However, when only unreduced vowel labels are considered, vowels in accented syllables were still more consistently transcribed than non-accented vowels, as shown on the right in Figure 8. Thus, a phrasal accent enhances the labeling consistency of the phones in an accented syllable.

6. Summary

In transcribing the spontaneous speech of the study, transcribers often disagreed about labels and their placement, but their disagreements were not without patterns. As in previous studies, disagreement was influenced by phonetic characteristics of the phones being labeled, lexical characteristics of the words containing the labeled segments, and prosodic properties of the speech.

There was almost never disagreement on whether a segment was a syllabic peak or not, but the non-peak consonant labels were more consistently applied than the peak vocalic labels. Within consonant label classes, resonant and glide labels were most likely to show disagreements involving labels of a different manner, and palatal and glottal labels were the most likely to show disagreements involving labels of a different place. In vocalic label selection, low front vowels and diphthongs had the highest proportion of disagreements outside of their class. However, individually the diphthongs were more consistently labeled than simple vowels. The simple vowels, in turn, were more consistently labeled than reduced vowels. In both the vocalic and consonant analyses, it was found that labels that indicated a narrower, more closely phonetic, transcription that indicated non-phonemic contrasts, were less consistently selected than labels for phonemes.

The greater consistency in labeling words than phones suggests that larger units are more consistently transcribed than smaller units, as might be expected. This tendency was seen within the two levels as well. Disagreements on word labels occurred on short, phonetically reduced words, and in the phonetic transcription there was more disagreement on the labeling of simple vowels than on the structurally longer diphthongs and vowels with off-glides. The effect of unit size is not the result of durational differences, because there was no effect of duration on labeling consistency.

Prominence also affected labeling consistency. Labels of phones in phrasally accented syllables were more consistently transcribed than phones in unaccented syllables. Phones in content words, which are more likely to received prominence, were also more consistently labeled than function word phones, although other differences between the two classes that were not examined in the study may also account for the difference, such as word class processing differences or word probabilities. In addition, word initial phones were more consistently labeled than non-initial phones, consistent with their greater perceptual salience.

The results indicate that, overall, linguistic annotation of the *Buckeye* corpus is consistent across transcribers. Consistency of annotation of the interview speech of the corpus was comparable to the consistency seen in other transcription projects of spontaneous speech. Unsurprisingly, annotation consistency was lower in this study than consistency rates reported for read speech, but factors influencing consistency in other studies were also seen to influence consistency in the current study. The information presented on where inconsistencies are likely to occur can guide future research using the annotated corpus data by indicating where annotation detail should be supplemented by direct acoustic analysis of the corresponding speech signal.

Acknowledgments

This project was conducted with support from NIDCD grant DC004330 to Mark Pitt, Keith Johnson, and Elizabeth Hume-O'Haire. Many thanks to Robin Dautricourt, Craig Hilts and Matthew Makashay for providing the transcriptions used in the study and for valuable discussion in developing the transcription guidelines. My special thanks to Craig Hilts for providing the ToBI transcription and for collaboration on consistency analysis.

References

Beckman, M.E. and Hirschberg, J. (1994) The ToBI annotation conventions. MS, The Ohio State University, Columbus, OH.

Bybee, J. and Hopper, P. (2001) *Frequency and the emergence of linguistic structure*. Philadelphia: John Benjamins.

Byrne, W., Finke, M., Khudanpur, S., McDonough, J., Nock, H., Riley, M., Saraclar, M., Wooters, C. and Zavaliagkos, G.. (1998) Pronunciation modeling using a hand-labelled corpus for conversational speech recognition. *Proceedings of ICASSP '98*.

Cucchiarini, C. (1996) Assessing transcription agreement: methodological aspects. *Clinical Linguistics and Phonetics*, 10(2), 131-155.

Deng, L. (1998) A dynamic, feature based approach to the interface between phonology and phonetics for speech modeling and recognition. *Speech Communication*, 24, 299-323.

Eisen, B. (1991) Reliability of speech segmentation and labeling at different levels of transcription. *Proceedings of Eurospeech-91*, 673-676.

Eisen, B., Tillman, H.G. and Draxler, C. (1992) Consistency of judgments in manual labeling of phonetic segments: The distinction between clear and unclear cases. *Proceeding of the 1992 International Congress of Speech and Language Processing*, 871-874.

Greenberg, S. (1998) Speaking in shorthand – A syllable-centric perspective for understanding pronunciation variation. *Proceedings of ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*, 47-56.

Greenberg, S., Hollenback, J. and Ellis, D. (1996) Insights into spoken language gleaned from phonetic transcription of the Switchboard corpus. *Proceeding of the 1996 International Congress of Speech and Language Processing*, 24-27.

Hayakawa, T., Kato H., Yoshinori, S. and Katsuhiko, S. (2002) Automatic phone segment alignment using statistical deviations from manual transcriptions. Paper presented at the First Pan-American/Iberian Meeting on Acoustics. Cancún, Mexico.

Jurafsky, D., Bell, A., Gregory, M. and Raymond, W.D. (2001) Probabilistic relations between words: Evidence from reduction in lexical production. In Bybee, J. and Hopper, P. (eds.) *Frequency and the emergence of linguistic structure*. Philadelphia: John Benamins (pp. 230-254).

Keating, P.A. (1997) Word-level phonetic variation in large speech corpora. In Alexiadou, A., Fuhrop, N., Kleinhenz, U. and Law, P. (eds.) (1998) *ZAS Papers in Linguistics*, 11, 35-50.

Keating, P.A., Byrd, D., Flemming, E. and Todaka, Y. (1994) Phonetic analyses of word and segment variation using the TIMIT corpus of American English. *Speech Communication*, 14, 131-142.

Kerswill, P. and Wright, S. (1991) The validity of phonetic transcription: Limitations of a sociolinguistic research tool. *Language Variation and Change*, 2, 255-275.

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

Pitrelli, J., Beckman, M.E. and Hirschberg, J. (1994) Evaluation of prosodic transcription labeling reliability in the ToBI framework. *Proceedings of the 1994 International Conference on Spoken Language Processing*, 123-126.

Pitt, M.A. and Samuel, A.G. (1995) Lexical and sublexical feedback in auditory word recognition. *Cognitive Psychology*, 29, 149-188.

Saraçlar, M. and Khudanpur, S. (2000) Pronunciation ambiguity vs pronunciation variability in speech recognition. *IEEE*, 1679-1682.

Shriberg, L.D. and Lof, G.L. (1991) Reliability studies in broad and narrow phonetic transcription. *Clinical Linguistics and Phonetics*, 5(3), 225:279.

Strik, H. and Cucchiarini, C. (1996) Modeling pronunciation variation for ASR: A survey of the literature. *Speech Communication*, 29, 225-246.

Wesenick, M.-B. and Kipp, A. (1996) Estimating the quality of phonetic transcriptions and segmentations of speech signals. *Proceedings of the 1996 International Conference on Spoken Language Processing*, 129-132.

Wightman, C. and Talkin, D. (1997) *The Aligner user's guide*. Washington D.C.: Entropic Research Laboratory, Inc.