SPONTANEOUS SPEECH, LAB SPEECH, AND EFFECTS ON INTONATION: SOME USEFUL FINDINGS FOR FIELDWORKERS (AND LABORATORY PHONOLOGISTS)*

Karsten A. Koch
University of British Columbia

This paper explores to what extent single sentence elicitations and scripted speech share acoustic properties with spontaneous utterances. Unlike previous studies of "lab speech" (eg. Klatt 1976, Lickley et al. 2005), the present analysis is of new data recorded in a fieldwork setting during research on Nɬeʔkepmxcin (Thompson River Salish). Results indicate that single sentence elicitations and spontaneous speech share many acoustic properties.

1. Background

Spontaneous speech does not always provide the utterances of theoretical interest to linguists, at least not with ample frequency. Instead of recording spontaneous speech, linguists have therefore resorted to more targeted techniques aimed at eliciting multiple instances of the structure of theoretical interest. These techniques may include single sentence elicitation (How do you say X? / Is X a good sentence?) (eg. Matthewson 2004), or tasks in which subjects read scripts in which the target utterances are frequently produced (eg. Lickley et al. 2005).

One issue is whether these alternate utterance types ("lab speech") are valid approximations to natural language situations (Thompson and Thompson 1992: 159 for a view against). The present paper examines this problem in terms of the acoustic properties of utterances collected in different speech contexts during fieldwork on Nɬeʔkepmxcin.[1] Nɬeʔkepmxcin (Thompson River Salish) is an Interior Salish language; the consultants for the present study are from the Lytton or *xʷqʷemcín* dialect. This study occurred as part of a larger project examining the properties of discourse phenomena such as focus, givenness and intonation in Nɬeʔkepmxcin (Koch 2008). While discourse phenomena occur in spontaneous speech, the utterances of specific interest to the theoretical linguist (for example, narrow focus versus wide focus utterances using specific NPs) might be quite rare. Thus, scripted speech and single sentence elicitations (types of "lab speech") were employed to elicit particular utterances.

* Many thanks to Flora Ehrhardt and Patricia McKay for their patience and teaching. This work has been supported by Jacobs and Kinkade Research Grants, NSERC Postgraduate Scholarships and a Strangway Fellowship to the author. Funding was also gratefully received through SSHRC Grants to Henry Davis, Lisa Matthewson and Hotze Rullmann.

[1] This leaves open the question whether utterances from different speech contexts differ in terms of syntax, morphology, semantics, and so on.

Actes du congrès annuel de l'Association canadienne de linguistique 2008.
Proceedings of the 2008 annual conference of the Canadian Linguistic Association.

Some previous research suggests that the acoustic differences between naturally produced speech and "lab speech" are less pronounced than the similarities (Klatt 1976: 1209, Lickley et al. 2005).

Lickley et al. (2005) had speakers read test sentences in an examination of postnuclear F0 minima in Dutch falling-rising questions. Of particular interest in this study, the authors investigated concerns that read speech in "laboratory phonology" studies is not a valid method for characterizing the intonation of spontaneous speech. On the other hand, even if there are "higher level" cognitive differences in the planning of read versus spontaneous speech (Levelt 1989), there may not be differences in "lower level" processes where "the planned utterance is translated into phonological/phonetic code .... That is, once a speaker has chosen a contour, it is a reasonable assumption that the contour's phonetic properties are largely or wholly predictable from phonetic and phonological factors alone" (2005:172). To test these hypotheses, the authors had 4 speakers produce a task-oriented dialogue using the Map Task (Anderson et al. 1991), which resulted in 21 questions directly comparable to those tested in the reading task. Both examination of the descriptive statistics and some statistical analysis failed to find any difference between read speech and spontaneous speech produced in the task-oriented experiment. The authors concluded that read speech can be "used as a source of evidence in experimental work that addresses phonological and phonetic questions" (2005:179), with its obvious practical advantages of using tightly controlled speech materials.

Similar findings in the present study would prove useful for field linguists investigating intonation, or other phonetic and phonological phenomena. It should be pointed out, however, that the type of data collected in the present study differs from typical read speech in standard laboratory phonology studies. This is due to the context of the present study, namely fieldwork on an endangered language with a purely oral tradition. Language consultants do not therefore read texts in their native language; because they are bilingual, they begin with an English script, and then develop an oral script in Nɬeʔkepmxcin. The scripted speech task thus involves both reading and translation, unlike typical "read speech" tasks. However, it is still a type of "lab speech."

A second type of data collected was single sentence elicitations (How do you say X? / Is X a good sentence?). This type of data again differs from read speech, though it also could be considered a type of "lab speech." Single sentence elicitation typically involves translation from English to the target language, and/or from the target language to English (see Mühlbauer 2008 on finer distinctions and more in-depth discussion of elicitations of this type).

Both of these data types were compared to spontaneous utterances in terms of their acousic properties.

2. Method

A corpus analysis was performed on Thompson Salish utterances of different modalities: spontaneous speech, single sentence elicitations, or scripted speech. The conversational data examined was part of a larger study on the expression

of the discourse notion of focus and givenness in Thompson River Salish (Koch 2008). The language data was collected from two female speakers of Nɬeʔkepmxcin in their late 60's (FE and PM). Both are speakers of the Lytton dialect, and fluently bilingual in English.

Recordings were made at the residence of either the language consultants or of the researcher, using a Marantz PMD 670, 671 or 660 digital audio recorder. Each consultant was recorded on a separate channel using a Countrymax Isomax EMW Lavalier microphone. The microphone was attached onto the exterior of the consultants' clothing, approximately at the sternum.

To account for declination effects, only utterances which were completed in a single breath group were considered for analysis. For each utterance, stressed lexical vowels at the left and right edges were identified in Praat (Boersma and Weenink 2007). Stressed vowels were measured for peak F0 (Hz), peak amplitude (db) and duration (ms), using automated scripts. Topline declination patterns for F0 and amplitude were calculated between these peaks (eg. t'Hart et al. 1990). Speech rate was also measured (syllables/sec). Where the Praat algorithm mismeasured F0 (eg. in the presence of glottalization, etc.), measurements were done by hand via visual inspection of the waveform, and automated measurements were disregarded. Figure 1 shows an example; see appendix B for a key to symbols and abbreviations.

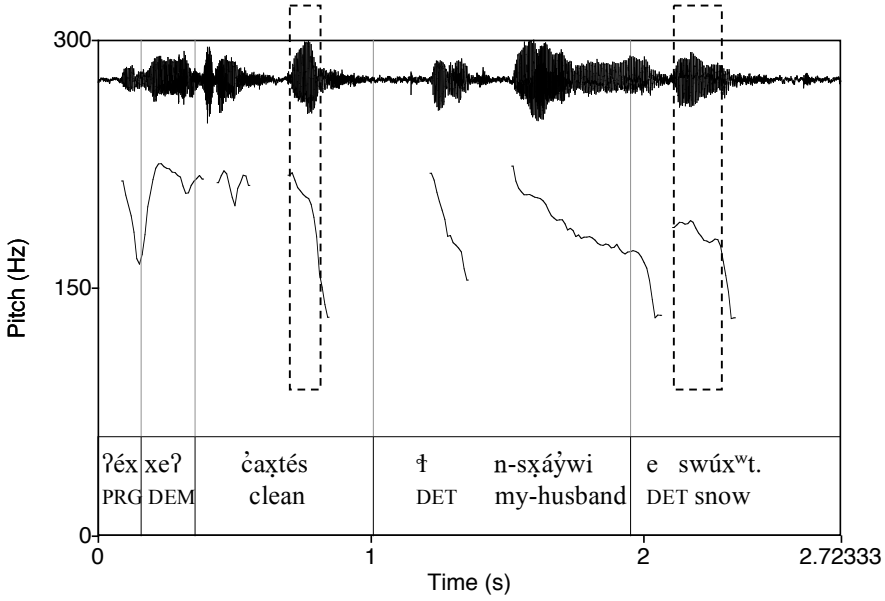


Figure 1. Spontaneous utterance: "My husband was cleaning up the snow." (Measured vowels are in the dashed boxes)

ANOVAs were used to test for main effects and interactions. ANOVAs for each variable were performed, with data source, focus type[2] and speaker as between factors. ANOVAs were conducted separately for left edge lexical stresses and right edge stressed vowels (to avoid violating the assumption of independence of measurements by including more than one measure from a single breath group). Separate ANOVAs were conducted for global utterance measures. Variables examined at the left and right edges were F0 maximum (Hz), amplitude maximum (db), and vowel duration (ms). Global utterance variables examined were speech rate (syllables/second), F0 declination rate between peaks (semitones/second), and amplitude declination rate between peaks (db/second). Pairwise comparisons of data source were performed as post-hoc tests, using Tukey's HSD to control for Type I error. Due to the number of ANOVAs conducted (9), a conservative p-value of 0.005 was adopted to control for experiment-wise error rate at a level of 0.05. To indicate trends in the data, I report three levels of significance in the tables used to illustrate results: $p<0.05$ and $p<0.01$ are marginally significant (indicated with * and ** respectively), while $p<0.005$ is the adjusted significance level (indicated by ***).

3. Results

Full statistical figures are reported in the accompanying tables in appendix A.

3.1 The Effects of Data Source: The Left Edge

For leftmost lexical vowels, only maximum amplitude showed a main, but marginal, effect of text ($F(2, 94)=3.77$, $*p<0.05$), with utterances from scripted conversation being louder than sentences from both spontaneous conversation or single sentence elicitation (figure 2). Post-hoc comparisons of amplitude peaks revealed that spontaneous utterances had significantly lower amplitude than scripted utterances (mean difference=–3.77 db, SD=1.13 db, $***p<0.005$), while single sentence elicitations were also marginally quieter than scripted conversational data (mean difference=–3.84 db, SD=1.39 db, $*p<0.05$). These results are illustrated in figure 2.

[2] Because the data was part of a larger study examining the expression of wide versus narrow focus in Thompson Salish (Koch 2008), focus type (wide focus or narrow focus) was an additional factor in this experiment. No interaction effects of focus type with data source were detected.

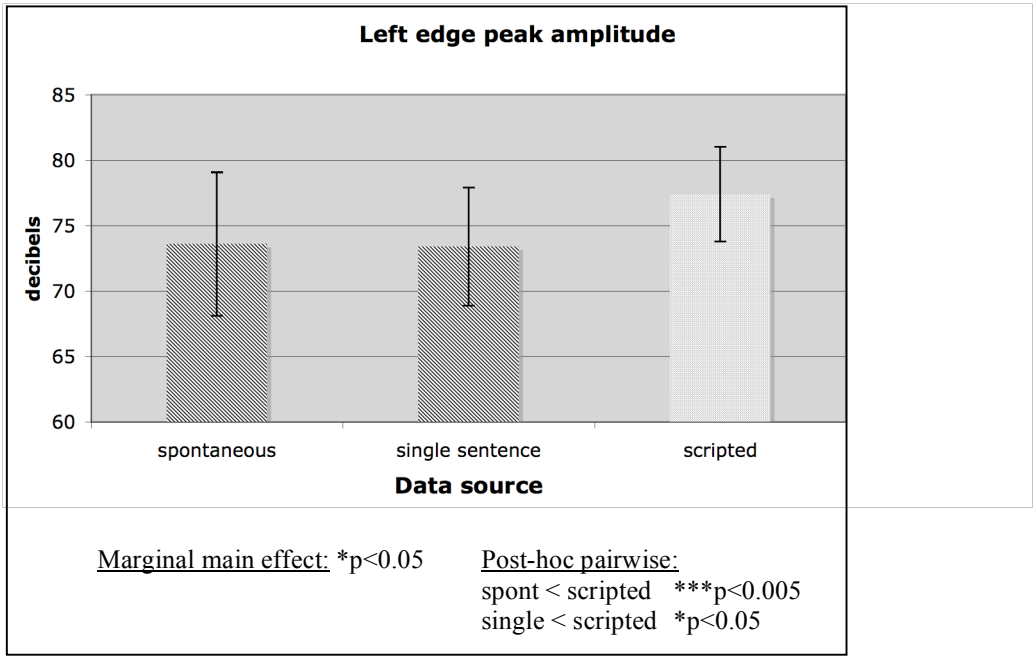


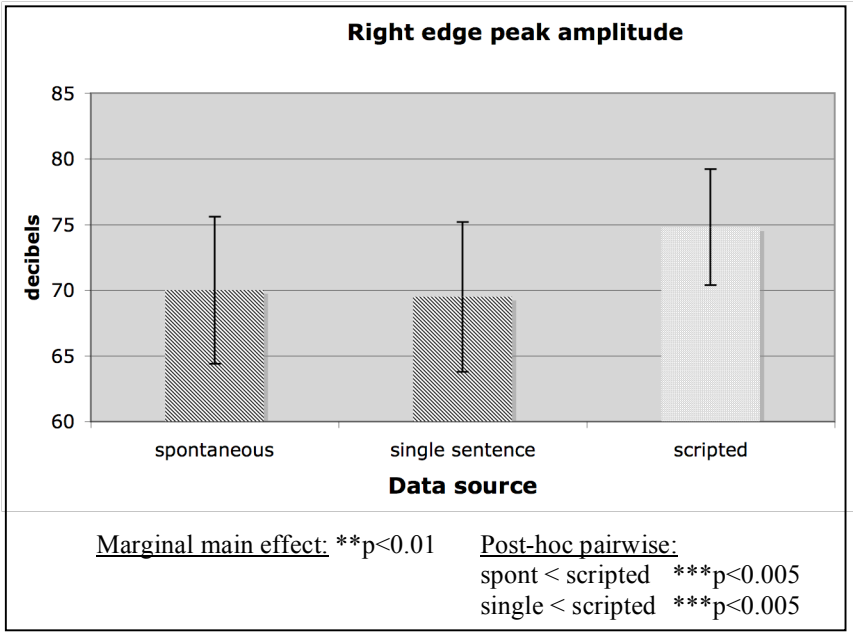
Figure 2. Left edge vowels are marginally louder in scripted utterances (error bars show one standard deviation from the mean)

As for F0 peaks, there was no main effect ($p>0.05$). Post-hoc pairwise comparisons detected only a weak effect: spontaneous conversation sentences had marginally lower F0 at the left edge than in scripted conversation (mean difference=9.17 Hz, SD=3.54 Hz, $*p<0.05$).

There was no effect of data source on vowel duration ($p>0.05$).

3.2 The Effects of Data Source: The Right Edge

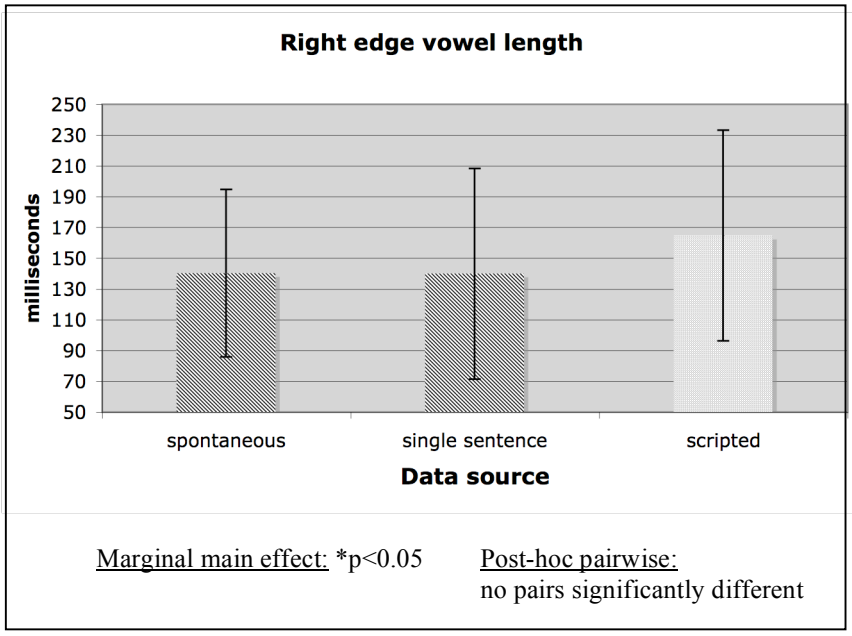
For rightmost stresses, there was again a marginal main effect of data source for amplitude ($F(2, 91)=5.074$, $**p<0.01$). As at the left edge, post-hoc comparisons showed that both spontaneous and single sentence elicitation utterances had significantly lower amplitude on rightmost stressed vowels than in scripted conversation. For spontaneous utterances, the mean difference was 4.75 db lower than scripted conversation (SD=1.10 db, $***p<0.005$). Single sentence elicitations had 5.25 db lower amplitude than scripted conversation (SD=1.31 db, $***p<0.005$). Figure 3 illustrates these results.



Marginal main effect: **p<0.01

Post-hoc pairwise:
spont < scripted ***p<0.005
single < scripted ***p<0.005

Figure 3. Right edge vowels are marginally louder in scripted utterances (error bars show one standard deviation from the mean)



Marginal main effect: *p<0.05

Post-hoc pairwise:
no pairs significantly different

Figure 4. Right edge vowels are marginally longer in scripted utterances (error bars show one standard deviation from the mean)

As illustrated in figure 4, vowel duration showed a marginally significant effect of data source ($F(2, 91)=3.316$, $^{*}p<0.05$), with final vowels in scripted conversation sentences tending to be longer (M=164.7 ms) than those in spontaneous (M=140.2 ms) or single sentence productions (M=139.9 ms).

There was no main effect of data source for F0 peaks on rightmost stressed vowels. However, the interaction of speaker and data source was significant ($F(2, 91)=6.196$, $^{***}p<0.005$). Post-hoc analysis for each speaker revealed that the source of the effect was a higher peak vowel F0 for FE in scripted conversation when compared to spontaneous conversation (mean difference=18.9 Hz, SD=5.1 Hz, $^{***}p<0.005$). PM tended to have *lower* F0 in scripted conversation sentences, but the effect was not significant.

3.3 The Effects of Data Source: Global Factors

There were no significant main effects of data source on utterance speech rate (syllables/second), utterance F0 declination rate, or amplitude declination rate. Post-hoc pairwise comparisons of data source also failed to detect any differences for these variables. While the mean F0 tended to be somewhat higher in scripted utterances, there was considerable variability among the three conditions (figure 5).

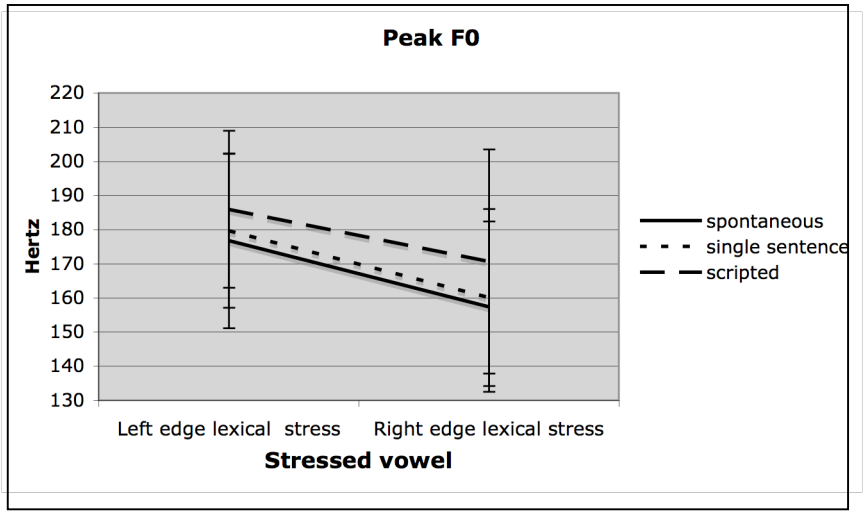


Figure 5. F0 declination does not differ significantly by data source (error bars show one standard deviation from the mean)

3.4 The Effects of Data Source: Summary

The primary differences in data source pit scripted conversation on the one hand against spontaneous conversation and single sentence elicitations on the other

hand. Scripted conversations tend to be louder, and have longer final stressed vowels. For FE scripted conversations have a higher stressed vowel F0 than spontaneous utterances. PM's F0 means trend in exactly the opposite direction, with scripted conversation tending to have lower F0 means (though not significantly so). Sentences from spontaneous conversations and single sentence elicitations thus appear to share many acoustic characteristics.

4. Discussion

Apart from a single marginal post-hoc effect (FE had marginally higher F0 on left edge lexical stresses in spontaneous conversation), spontaneous utterances did not differ significantly from single sentence elicitation. These results suggest that, given a sufficiently detailed context (Matthewson 2004), speakers are able to deliver single sentence utterances whose intonation closely approximates the speech melody used in spontaneous conversation. For researchers who want to study aspects of intonation without collecting large samples of spontaneous conversation (or for those working with a single speaker of an endangered language, where such an endeavour is not possible), this is an important result. Carefully elicited sentences could form the core of an analysis of intonational properties of a language, perhaps to be checked with samples of spontaneous conversation after basic principles have been predicted.

Utterances taken from scripted conversation (role-playing dialogues), on the other hand, were found to differ primarily in being louder, having longer duration on rightmost stressed vowels, and for FE, higher F0 on stressed vowels. These results suggest that scripted role-playing conversation generates less natural-sounding intonation.

However, let us consider the finding of greater amplitude in scripted conversation utterances in more detail. Since, in the collection of data, microphone recording levels were reset for each session, and lapel microphones were attached in slightly different positions each time, it is possible that the observed amplitude difference is due to variability in recording conditions. However, since spontaneous, scripted and single sentence recordings were all carried out across many recording sessions, it is expected that this variation in recording conditions is relatively evenly distributed among the utterances comprising the data set. A second possibility is that speaker orientation is more directly oriented towards the lapel microphone during scripted conversation. Since the speakers are working from an English text printed on a piece of paper, they are spending more time looking downward in front of them (towards where the lapel microphone is attached, approximately at the sternum). This orientation in speech may lead to higher amplitude in the recordings for scripted conversation. In spontaneous utterances, or single sentence elicitations, no written material is involved, so speakers are free to look elsewhere, and speech is less directly aimed at the lapel microphone. If speaker orientation accounts for amplitude differences, then utterances generated from scripted conversation may be more similar to spontaneous conversation than the results suggest (and in line with findings by Klatt 1976, Lickley et al. 2005, for "lab speech").

5. Conclusion

In this case study, I tested the effect of data source (spontaneous conversation, scripted conversation, or single sentence utterances) on sentence intonation. Findings suggest that speakers provide comparable intonation in both single sentence elicitations and spontaneous utterances. Scripted conversation, on the other hand, differs primarily in greater amplitude, and longer duration of rightmost stressed vowels.

To the extent that the present findings can be replicated elsewhere, they represent an important finding for linguists seeking to investigate acoustic properties of speech in endangered (and other) languages. The ability to examine acoustic properties representative of spontaneous speech by engaging in carefully controlled single sentence elicitation can save much research time.

References

Anderson, A., M. Bader, E.G. Bard, E. Boyle, G.M. Doherty, S. Garrod, D. Isard, J. Kowtko, J. McAllister, J. Miller, C. Sotillo, H.S. Thompson, and R. Weinert. 1991. The HCRC Map Task Corpus. *Language and Speech* 34: 351-366.

Boersma, Paul, and David Weenink. 2007. Praat: doing phonetics by computer (Version 4.5.14) [Computer program]. Downloaded from http://www.praat.org.

Klatt, D.H. 1976. Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *Journal of the Acoustical Society of America* 59: 1208-1221.

Koch, Karsten A. 2008. Intonation and focus in Nɬeʔkepmxcin (Thompson River Salish). Doctoral dissertation, University of British Columbia.

Kroeber, Paul. 1997. Relativization in Thompson Salish. *Anthropological Linguistics* 39(3): 376-422.

Levelt, W. J. M. 1989. Speaking: From intention to articulation. Cambridge, MA: MIT Press.

Lickley, Robin J., Astrid Schepman and D. Robert Ladd. 2005. Alignment of "phrase accent" lows in Dutch falling-rising questions. *Language and Speech* 48(2): 157-183.

Matthewson, Lisa. 2004. On the methodology of semantic fieldwork. *International Journal of American Linguistics* 70:369-415.

Muehlbauer, Jeff. 2008. kâ-yôskâtahk ôma nêhiyawêwin: The representation of intentionality in Plains Cree. Doctoral dissertation, University of British Columbia.

't Hart, Johan, René Collier, and Antonie Cohen. 1990. *A Perceptual Study of Intonation.* Cambridge: Cambridge University Press.

Thompson, L.C., and M.T. Thompson. 1992. *The Thompson Language*. Missoula: UMOPL 8.

Appendix A: Statistical Figures

Table 1. Main effects of data source: the leftmost lexical stress

Acoustic variable	Data Source	Mean (SD)	n	F	p	df
Max F0 (Hz)	spontaneous	176.7 (25.6)	54	0.998	0.373	2/94
	scripted	185.9 (23.0)	30			
	single sentence	179.6 (22.5)	22			
Max amplitude (db)	spontaneous	73.6 (5.5)	54	3.77	*0.027	2/94
	scripted	77.4 (3.6)	30			
	single sentence	73.4 (4.5)	22			
Vowel duration (ms)	spontaneous	116.2 (54.3)	54	0.878	0.419	2/94
	scripted	126.9 (38.1)	30			
	single sentence	124.4 (35.9)	22			

Table 2. Significant post-hoc pairwise comparisons of data source for leftmost stress (adjusted with Tukey's HSD)

Acoustic variable	Significant pairs	Mean diff. (SD)	p	95% Conf. Int. Lower	Upper
Max F0 (Hz)	spont. < script	-9.17 (3.54)	*0.03	-17.60	-0.74
Max amp (db)	spont. < script	-3.77 (1.13)	***0.003	-6.45	-1.09
	single < script	-3.94 (1.39)	*0.015	-7.25	-0.64
V dur. (ms)	none				

Key: SD=standard deviation, n=# of observations, p=probability, df=degrees of freedom, Conf. Int. = confidence interval, *=sig. at $p<0.05$, **=sig. at $p<0.01$, ***=sig. at $p<0.005$ [note: *** is controlled for a family-wise error rate of 0.05]

Table 3. Main effects of data source: the rightmost lexical stress

Acoustic variable	Data Source	Mean (SD)	n	F	p	df
Max F0 (Hz)	spontaneous	157.4 (24.9)	49	0.843	0.434	2/91
	scripted	170.6 (32.8)	31			
	single sentence	160.1 (25.9)	23			
Max amplitude (db)	spontaneous	70.0 (5.6)	49	5.074	**.008	2/91
	scripted	74.8 (4.4)	31			
	single sentence	69.5 (5.7)	23			
Vowel duration (ms)	spontaneous	140.2 (54.4)	49	3.316	*0.041	2/91
	scripted	164.7 (68.4)	31			
	single sentence	139.9 (68.4)	23			

Table 4. Significant post-hoc pairwise comparisons of data source for righmost stress (adjusted with Tukey's HSD)

Acoustic variable	Significant pairs	Mean diff. (SD)	p	95% Conf. Int. Lower	Upper
Max F0 (Hz)	spont.<script	-13.21 (3.96)	***0.004	-22.65	-3.77
Max amp (db)	spont. < script	-4.75 (1.10)	***<.001	-7.36	-2.14
	single < script	-5.25 (1.31)	***<.001	-8.38	-2.12
V dur. (ms)	none				

Key: SD=standard deviation, n=# of observations, p=probability, df=degrees of freedom, Conf. Int. = confidence interval, *=sig. at p<0.05, **=sig. at p<0.01, ***=sig. at p<0.005 [note: *** is controlled for a family-wise error rate of 0.05]

Table 5. Interaction effect of data source by speaker: F0 maximum (Hz) on the rightmost stress

Acoustic variable	Data Source	Mean (SD)	n	F	p	df
Max F0	Speaker*data source interaction.....			6.196	***0.003	2/91
Max F0 (FE)	spontaneous	174.1 (19.1)	22			
	scripted	193.0 (7.56)	20			
	single sentence	175.6 (22.2)	13			
Max F0 (PM)	spontaneous	143.7 (20.4)	27			
	scripted	129.8 (16.9)	11			
	single sentence	139.9 (13.5)	10			

Table 6. Significant post-hoc pairwise comparisons of data source on F0 peak, by speaker

Acoustic variable	Significant pairs	Mean diff. (SD)	p	95% Conf. Int. Lower	Upper
Max F0 (FE)	spont. < script	-18.9 (5.1)	***0.001	-31.11	-6.65

Key: SD=standard deviation, n=# of observations, p=probability, df=degrees of freedom, Conf. Int. = confidence interval, *=sig. at $p<0.05$, **=sig. at $p<0.01$, ***=sig. at $p<0.005$ [note: *** is controlled for a family-wise error rate of 0.05]

Table 7. Main effects of data source: global utterance characteristics (including speaker means)

Acoustic variable	Data Source	Mean (SD)	n	F	p	df
Speech rate	spontaneous	4.00 (1.02)	49	0.27	0.764	2/91
(syllables/sec)	scripted	3.70 (0.68)	31			
	single sentence	3.93 (0.90)	23			
FE only	spontaneous	3.64 (0.74)	22			
	scripted	3.66 (0.56)	20			
	single sentence	3.96 (0.76)	13			
PM only	spontaneous	4.28 (1.14)	27			
	scripted	3.78 (0.89)	11			
	single sentence	3.89 (1.10)	10			
F0 declination	spontaneous	1.47 (1.21)	49	1.462	0.237	2/91
(semitone/sec)	scripted	1.41 (1.32)	31			
	single sentence	1.05 (0.78)	23			
FE only	spontaneous	1.34 (0.89)	22			
	scripted	0.96 (0.58)	20			
	single sentence	0.95 (0.76)	13			
PM only	spontaneous	1.58 (1.42)	27			
	scripted	2.22 (1.85)	11			
	single sentence	1.19 (0.82)	10			
Amplitude	spontaneous	1.14 (2.75)	49	0.405	0.668	2/91
declination	scripted	0.90 (1.52)	31			
(db/sec)	single sentence	1.45 (2.81)	23			
FE only	spontaneous	0.90 (2.88)	22			
	scripted	0.51 (0.89	20			
	single sentence	1.12 (2.99)	13			
PM only	spontaneous	1.34 (2.67)	27			
	scripted	1.60 (2.13)	11			
	single sentence	1.88 (2.66)	10			

Appendix B: Symbols and Abbreviations

For reasons of space and clarity, I do not provide full morphological breakdowns for all nouns, adjectives, adverbs, and so on. Abbreviations used in the gloss (based on Thompson and Thompson 1992, Kroeber 1997) are as follows:

DEM = demonstrative
D, DET = determiner
PROG, PRG = progressive

Data are presented in the orthography developed in Thompson and Thompson (1992), and Kroeber (1997). The phonemic key to the *orthography* is as follows; symbols not listed have the standard IPA interpretation:

c = [tʃ] or [č]
ç = [ts]
ć = [ts’]
e = [e, æ, a, ɛ, ə]
ẹ = [ʌ]
i = [i, ei, ai]
o = [o, ɔ]
s = [ʃ] or [š]
ṣ = [s]
u = [u, o, ɔ]
x̣ = [χ]
y = [y, i].

See Thompson and Thompson (1992) in particular for the phonetic realizations of phonemic vowels across contexts. Nɬeʔkepmxcin [z] is more lateral than English [z], though there may be considerable regional or speaker variation.